دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

دانشکده مهندسی برق

پایان نامه کارشناسی ارشد

عنوان پایان نامه:

بهبود عملکرد روش HMM در دیکدر ATP گفتار پیوسته فارسی

نگارش : ساناز علیزاده

استاد راهنما : آقای دکترصیادیان

بهمن ۱۳۸۵

بسمه تعالي

تاريخ : .....
شماره : .....

فرم اطلاعات پايان نامه
كارشناسي ارشد و دكترا

دانشگاه صنعتي اميركبير
(پلي تكنيک تهران)

معاونت پژوهشي
فرم پروژه تحصيلات تكميلي ۷

مشخصات دانشجو

نام و نام خانوادگي : ساناز عليزاده ✓ دانشجوي آزاد ☐ بورسيه ☐ معادل

شماره دانشجويي: ۸۳۱۲۳۱۱۸ دانشكده : برق رشته تحصيلي: مخابرات سيستم

نام و نام خانودگي استاد راهنما : دكتر ابولقاسم صيادیان

عنوان به فارسي: بهبود عملكرد روش HMM در ديكدر ATP گفتار پيوسته فارسی

عنوان به انگليسي: Improvement of HMM performance in Persian ATP for continuous speech

نوع پروژه: كارشناسي ارشد دكترا ✓ **كاربردي** ☐ **بنيادي** ✓ **توسعه اي** ☐ **نظري**

تاريخ شروع: ۲۷ تير ۱۳۸۳ **تاريخ خاتمه : ۱۵ بهمن ۱۳۸۵** **تعداد واحد: ۶**

سازمان تامين كننده اعتبار : مركز تحقيقات مخابرات ايران

واژه هاي كليدي به فارسي : بازشناسی گفتار پيوسته - مبدل سيگنال شنيداری به معادل آن

واژه هاي كليدي به انگليسي : HMM – Continuous speech recognition - ATP

نظرها و پيشنهادها به منظور بهبود فعاليت هاي پژوهشي دانشگاه:

استاد راهنما:

دانشجو:

امضاء استاد راهنما : **تاريخ:**

نسخه ۱: معاونت پژوهشي
نسخه ۲ : كتابخانه و به انضمام دو جلد پايان نامه به منظور تسويه حساب با كتابخانه و مركز اسنادومدارک علمي

چکیده پایان نامه :

تغییرات مشخصه آکوستیکی واج ها تحت متن های مختلف موجب شده است که در پیاده سازی سیستم‌های بازشناسی گفتار، از واحد های گفتاری وابسته به متن مانند هجا و نیم هجا که اثرات آواهای مجاور را نیز در نظر می‌گیرند، استفاده شود. با توجه به اینکه زبان فارسی از دسته زبانهایی می‌باشد که دارای ساختار هجایی ساده‌ای است، در این تحقیق واحد گفتاری نیم هجا برای مدلسازی طیفی مورد توجه قرار گرفته است و آزمون های متعددی برای تصدیق مطلب فوق در طی انجام پروژه صورت گرفته است. به علت فقدان پایگاه داده فارسی مبتنی بر نیم هجا، تلاش زیادی جهت طراحی متن و جملات مورد نیاز برای پایگاه داده در طی این تحقیق انجام پذیرفته است و داده‌های گفتاری مربوط به دو گوینده زن و دو گوینده مرد برای ارزیابی مدلها بیان و ضبط شده و به صورت با سرپرستی در سطح واکه و نیم هجا برچسب زده شده است.

با توجه به اینکه تشخیص نیم هجاها، ابتدا از آشکارسازی سکوت و واکه آغاز می‌گردد، در اولین قدم تمام توجه ما به آشکارسازی واکه ها معطوف شده است. در بازشناسی واکه ها، از ترکیب مدل آماری مارکوف پنهان با پارامترهای آکوستیکی مانند انرژی میانگذر استفاده شده است. در این پروژه محدوده واکه ها با بهره مندی از ویژگی های مدل آکوستیکی مانند سادگی، سرعت و ناوابسته بودن آن به گوینده ها، مشخص شده است. سپس با ترکیب نتایج حاصل از پارامترهای آکوستیکی و مدل آماری مارکوف پنهان به نتایج بسیار مناسبی در بازشناسی واکه ها دست یافتیم. در این پروژه در بهترین حالت ها، در گفتار پیوسته به خطای ۸/۹۸٪ و در گفتار گسسته به خطای ۲/۸۷٪ دست یافتیم.

Abstract

فهرست مطالب

فصل اول- مقدمه

فصل دوم-روش های رایج در بازشناسی گفتار

فصل سوم- پایگاه داده

فصل چهارم-پیاده سازی مدل پیشنهادی

۱-۱ پیش گفتار

سیستمهای بازشناسی گفتار[1]، از پرکاربردترین سیستمهای کاربردی جهت ارتباط طبیعی انسان و ماشین می‌باشند. اگرچه بسیاری از سیستم های بازشناسی گفتار در کاربردهای محدود به موفقیتهایی دست یافته اند، اما با توجه به اهمیت و کارآیی این سیستم ها، تلاش برای بهبود و توسعه محدوده عملکرد آنها ادامه دارد. بازشناسی گفتار، فرآیند تبدیل صحبت به متن و بطورکلی تر تبدیل صحبت به رشته ای از برچسب ها می‌باشد. با توجه به تعریف بازشناسی گفتار، در نگاه اول بازشناسی گفتار فرآیندی ساده به نظر می‌رسد و دو مرحله برای اجرای آن در ذهن تداعی می‌شود:

- تصمیم گیری در مورد نوع آوای تلفظ شده
- جستجوی کلمه حاصل از توالی آواهای شناخته شده، در فرهنگ لغاتی که از کلمات بیان شده تهیه شده است.

سیستم های بازشناسی گفتار که جهت بازشناسی ارقام و واکه‌ها در دهه ۵۰ طراحی شده بودند، بازده چشمگیری داشتند[۱]. بازشناسی در این سیستم ها بر اساس دو مرحله عنوان شده در بالا انجام می‌شد. به نظر می‌آمد بتوان با تعمیم تکنیکهای به کاررفته در این سیستم‌ها، این نتایج را برای سیستم‌های پیچیده‌تری تعمیم

[1] Speech recognition

داد. اما متاسفانه کاربرد تکنیک های استفاده شده در سیستم های با عملکرد و کاربردهای محدود نتوانست نتایج و کارآیی مناسبی در سیستمهای پیچیده‌تر ارائه دهد.

سیستم‌هایی بازشناسی گفتار با قابلیت‌های بهتر دارای پارامترهای بیشتری نسبت به سیستم‌های ساده می-باشند. دسته‌بندی این سیستم ها بر اساس پارامترهای گفتار ورودی انجام می‌شود و تغییر هر یک از آنها در گفتار ورودی موجب پیچیده شدن فرآیند بازشناسی گفتار می‌شود. تنوع در این فاکتورها می تواند ناشی از تعدد و تنوع گوینده‌ها، تنوع در حالت های ادای گفتار، تنوع در لهجه ها و تنوع در محیط پیرامون و یا مربوط به گستردگی دامنه لغات باشد. در حال حاضر هیچ روشی وجود ندارد که بتواند تنوع در تمامی منابع این پارامترها را پوشش دهد. سیستم های طراحی شده با هدف پوشش برخی از این عوامل و برای کاربردهای خاص طراحی شده‌اند. در بخش بعدی، ۵ عامل اصلی تنوع در گفتار ورودی که موجب پیچیدگی بازشناسی گفتار می‌شوند را بررسی خواهیم نمود.

۱-۲ مسائل مطرح در سیستم های بازشناسی گفتار

مهمترین مساله‌ای که در طراحی یک سیستم بازشناسی گفتار باید در نظر گرفته شود، پیوسته و یا گسسته بودن گفتار ورودی می‌باشد. گفتار گسسته[1] به گفتاری گفته می‌شود که در آن گوینده کلمات را بصورت مجزا و با فاصله زمانی مشخصی بیان می‌کند، در واقع گوینده پس از بیان هر کلمه مکث می‌کند. در حالیکه در گفتار پیوسته[2] فاصله زمانی معین و از پیش تعیین شده‌ای بین کلمات متوالی وجود ندارد و گفتار بصورت رشته‌ای از لغات متوالی بیان می‌شود.

یکی دیگر از پارامترهای تعیین کننده در سیستم بازشناسی گفتار، دامنه لغات قابل بازشناسی توسط این سیستم می باشد[۲]. پیچیدگی یک سیستم بازشناسی گفتار پیوسته و یا گسسته با پایگاه داده‌ای با تعداد لغات محدود، بسیار کمتر از پیچیدگی یک سیستم بازشناسی گفتار پیوسته و یا گسسته با پایگاه داده‌ای با دایره لغات گسترده می باشد. هنگامیکه دادگان پایگاه داده محدود است، می توان کلمه را به عنوان واحد گفتاری در نظر

[1] discrete
[2] continuous

گرفت و برای هر کلمه یک مدل آکوستیکی طراحی نمود. اما با افزایش دایره لغات، طراحی مدل برای هر یک از آنها و نیز جستجو در چنین پایگاه داده‌ای، بسیار وقت گیر می‌باشد. اکثر سیستم های تجاری موجود دارای دامنه لغات محدود می‌باشند. اما نیاز به سیستم‌های با دایره لغات وسیع تر موجب شده است که مدلهای گفتاری کوچکتری طراحی شوند. در این نوع سیستم‌ها، مدلها از واحدهای زیر لغوی[۱] به جای کلمه برای مدلسازی گفتاری بهره می‌برند. سیستم های بازشناسی مبتنی بر واج[۲]، هجا[۳] و نیم هجا[۴] نمونه هایی از این سیستم‌ها هستند.

همچنین در سیستم های بازشناسی گفتار پیوسته از مدلهای زبانی[۵] یا دستور زبان ساختگی برای محدود ساختن تعداد جملات قابل طراحی با کلمات پایگاه داده استفاده می‌شود. ساده‌ترین نوع مدل زبانی مدلی است که ترتیب خاصی از کلمات را مجاز می‌سازد و نوع رایج تر مدلی است که تقریباً مشابه دستور زبان طبیعی می‌باشد و بصورت قواعد وابسته به متن می‌باشد. ترکیب تاثیر دامنه لغات و دستور زبان می تواند با سرگشتگی[۶] اندازه گیری می‌شود[۱]. سرگشتگی بیانگر تعداد متوسط لغاتی است که می توانند در هر نقطه تصمیم گیری رخ دهد که با افزایش پیچیدگی سیستم بازشناسی گفتار افزایش می‌یابد.

تفاوتهای گویندگان از نظر جنسیت، سن و آهنگ و لحن گفتار نیز موجب تنوع در سیگنال آکوستیکی می-شود. سیستم های بازشناسی وابسته به گوینده[۷] تنها برای یک یا تعداد محدودی گوینده خاص طراحی و آموزش داده می شوند، درحالیکه سیستم های ناوابسته به گوینده[۸] فرآیند بازشناسی گفتار را مستقل از گوینده و برای تمامی گویندگان انجام می‌دهند. سیستم های نوع اول تنها قابلیت ایجاد تنوع در لحن و نرخ گفتار گوینده‌ای که با صدایش آموزش داده شده‌اند را دارند. در حالیکه سیستم های نوع دوم می توانند تنوع سنی، جنسی و حالتهای گفتار گوینده‌های متفاوت را پوشش دهند.

---

[1] Sub-word
[2] phonem
[3] syllable
[4] Demi-syllable
[5] Language modeling
[6] perplexity
[7] Speaker dependent
[8] Speaker independent

عامل مهم دیگری که سبب ایجاد تغییر و تنوع در سیگنال گفتار ورودی و در نتیجه تضعیف عملکرد سیستم بازشناسی گفتار می‌شود، نویز می‌باشد. نویز می‌تواند ناشی از عوامل محیطی همچون ترافیک، باران و یا سر و صدای محیط پیرامون و یا ناشی از سرفه و یا بازدم گوینده باشد. نویزهای دیگری نیز که در بخش های مختلف سیستم بازشناسی گفتار به وجود می آیند موجب کاهش کیفیت سیستم می‌شوند.

۱-۳ مقایسه سیستم بازشناسی گفتار پیوسته و گسسته

هر چند جدا سازی کلمات در سیستم های بازشناسی گفتار گسسته به دلیل وجود وقفه بین کلمات به سادگی انجام می‌شود، اما مکث گوینده پس از بیان هر کلمه موجب می‌شود گفتار از حالت طبیعی خارج شود. بازشناسی گفتار پیوسته به سه دلیل بسیار مشکل تر از بازشناسی گفتار گسسته می‌باشد:

- نبودن فاصله زمانی معین بین کلمات
- تاثیر آواهای مجاور بر یکدیگر
- تغییر لحن و تلفظ کلمات به علت قرار گرفتن در ساختار جمله

از آنجاییکه در گفتار پیوسته فاصله زمانی معینی بین کلمات وجود ندارد، اولین مشکل، تعیین مرز ابتدایی و انتهایی کلمات به کار رفته در یک جمله می‌باشد تا با توجه به مرز تعیین شده بتوان نوع آنها را تعیین نمود.

مشکل دیگر ناشی از تاثیر آواهای مجاور در کلمات متوالی بر هم می باشد که به عنوان اثرات هم تولیدی[1] شناخته می‌شود و دارای ساختاری پیچیده است. در گفتار پیوسته تاثیر آواهای مجاور بر یکدیگر در درون کلمه و در بین کلمات متوالی بصورت تاثیر آوای انتهایی یک کلمه بر آوای ابتدایی کلمه مجاورش مشاهده می‌شود. اثرات هم تولیدی در گفتار موجب می‌شود تلفظ یک کلمه با توجه به موقعیت آن کلمه در جمله نسبت به کلمات مشابه دیگر تغییر کند. چنانچه در یک سیستم بازشناسی گفتار، تاثیر آواهای متوالی بر یکدیگر مدلسازی شود به این

[1] co-articulation

سیستم یک سیستم بازشناسی گفتار وابسته به متن[1] گفته می‌شود و در غیر این صورت یک سیستم بازشناسی گفتار غیر وابسته به متن[2] می‌باشد.

وقتی کلمات در یک جمله به دنبال یکدیگر بیان می شوند، لحن و نوع تلفظ هر یک از کلمات نسبت به حالتیکه بصورت مجزا بیان می‌شوند تغییر می‌کند. همچنین در نوعی از گفتار پیوسته که به آن گفتار بی تکلف[3] گفته می‌شود و در واقع همان گفتار رایج در میان مردم می‌باشد، ممکن است نوعی اختصار و یا حذف کلمات قرینه به اقتضای توالی آنها در جملات رخ دهد. اثرات دیگر ناشی از سرعت گفتار نیز در این بین خودنمایی می-کنند که سبب ادغام آواها و تولید آوایی بینابین می‌شود. لازم به ذکر است که این نوع گفتار از قوانین دستوری پیروی نمی‌کند و به همین دلیل بازشناسی این نوع گفتار پیوسته نسبت به گفتار پیوسته نوشتاری که دارای قواعد دستوری مشخص می‌باشد، بسیار پیچیده‌تر است.

۱-۴ طراحی یک سیستم بازشناسی گفتار

سیستم های بازشناسی گفتار اولیه، از روش انطباق با مدل استفاده می‌کردند. همانطور که از نام این روش برمی‌آید، در این روش ابتدا یک قالب[4] برای هر یک از لغات پایگاه داده طراحی می‌شود. سپس ورودی با کلیه قالب ها مقایسه می‌شود و با توجه به میزان شباهت ورودی با قالب ها، بهترین تطابق، معادل ورودی را مشخص می‌نماید. لازم به ذکر است که یک قالب می‌تواند از میانگین گیری بر روی نمونه های مختلف ضبط شده از تلفظ های متعدد یک کلمه تهیه شود. این روش برای بازشناسی گفتار گسسته با دایره لغات محدود دارای کارآیی مناسبی می‌باشد.

اما متاسفانه تعمیم این روش به گفتار پیوسته و گفتار گسسته با پایگاه داده بزرگ به دلیل مشکلات ناشی از تهیه قالب برای این تعداد کلمات و نیز مشکل تعیین مرز بین کلمات امکان پذیر نیست. از آنجاییکه ارتباط بین

[1] Context dependent
[2] Context independent
[3] spontaneously
[4] prototype

کلمات موجود در پایگاه داده و کلمات به کار رفته در یک جمله از یک رابطه قطعی و یقینی پیروی نمی‌کند، باید از روشهای ریاضی به عنوان جایگزین قالب ها استفاده نمود.

در یک سیستم بازشناسی گفتار پیوسته، گفتار بیان شده از طریق میکروفن پس از نمونه برداری با فرکانس ۸ یا ۱۶ کیلو هرتز بصورت یک سیگنال دیجیتال به واحد پیش پردازش وارد می‌شود. سپس بردارهای ویژگی این نمونه های فریم بندی شده استخراج می‌گردند. بنابراین سیگنال آکوستیکی حاصل از بیان جمله W، بصورت رشته ای از مشاهدات آکوستیکی درمی‌آید. نوع بردار مشاهده آکوستیکی یا همان بردار ویژگی[1]، بسته به تکنیک به کار رفته در استخراج بردارهای ویژگی می‌تواند از نوع MFCC , LPC و غیره باشد. در نهایت مدل آکوستیکی سیگنال ورودی بصورت زیر خواهد بود:

$$A = a_1, a_2, \ldots, a_m \qquad a_i \in \mathrm{A} \qquad (۱-۱)$$

یک جمله شامل یک رشته از n کلمه و یا بطور کلی تر شامل n واحد آکوستیکی مانند کلمه و یا زیر کلمه می‌باشد. در یک سیستم بازشناسی، پایگاه داده υ کلیه واحدهای آکوستیکی را که هر یک از دنباله ای از بردارهای ویژگی تشکیل شده‌اند، پوشش می‌دهد. w_i بیانگر سمبل ورودی در لحظه i می‌باشد.

$$\mathbf{W}` = w_1, w_2, \ldots, w_n \qquad w_i \in \upsilon \qquad (۲-۱)$$

وظیفه یک سیستم بازشناسی گفتار، پیداکردن محتمل ترین دنباله لغات و یا سمبل‌ها از روی دنباله مشاهدات آکوستیکی می‌باشد با کمک رابطه بیز می توان $P(W \mid A)$ رابصورت زیر نوشت:

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

احتمال مشاهده دنباله آکوستیکی $P(A)$، برای یک دنباله مشاهده مشخص به ازای تمام W های ممکن، مقداری ثابت می باشد. بنابراین $P(A)$ را در محاسبات مربوط به بیشینه سازی $P(W \mid A)$ وارد نمی‌کنیم. لذا خواهیم داشت:

[1] Feature vector

$$\hat{W} = \arg\max_w P(W)P(A|W) \qquad (۴-۱)$$

حساس‌ترین بخش در بازشناسی گفتار، تعیین احتمال $P(A|W)$ می‌باشد. برای محاسبه این احتمال از روشهای آماری متعددی استفاده می‌گردد. برای محاسبه $P(W)$ نیاز به یک مدل زبانی داریم که با توجه به ماهیت گفتار، مدلی آماری می‌باشد. در نهایت برای پیداکردن محتمل ترین دنباله سمبل ها از رابطه (۱-۴) ، به یک الگوریتم جستجو نیاز داریم. لذا از روشهای ریاضی برای محاسبه $P(A|W)$ استفاده می شود. تمامی روشهای رایج برای محاسبه این احتمال به جز روش پیچش زمانی پویا[1] روشهای آماری می‌باشند. در کلیه روشهای آماری، مستقل از مدل آکوستیکی انتخاب شده برای بازشناسی گفتار، هر واحد آکوستیکی با یک مدل آماری مدلسازی می‌شود.

با توجه به اینکه پارامترهای آماری یک واحد آکوستیکی در طول این واحد تغییر می‌کند و ثابت نمی‌ماند، برای مدلسازی آماری هر یک از این واحدها، آن را به قطعه[2] هایی تقسیم می‌کنند، به گونه ایکه بتوان توزیع آماری عناصر آن قطعه را ثابت در نظر گرفت. انتخاب مدل آماری مناسب از چهار جهت حائز اهمیت می‌باشد که عبارتند از:

- مدلسازی توالی قطعات
- نمایش همبستگی بردارهای یک قطعه
- نمایش همبستگی بین قطعات وتاثیر قطعات متوالی بر یکدیگر
- در بر گیرندگی تنوع بردارهای ویژگی و تغییرات مشخصات آماری آنها در طول زمان

همچنین انتخاب مناسب طول قطعه به گونه ایکه مدل آماری منتخب بتواند تنوع در مشخصات آماری و پارامترهای یک سیستم بازشناسی گفتار مانند تنوع در گوینده، لهجه و حالات مختلف گویش را پوشش دهد نیز

[1] DTW
[2] Segment

حائز اهمیت می‌باشد. رایج ترین مدل آماری مورد استفاده در بازشناسی گفتار مدل مارکوف پنهان[1] (HMM) می‌باشد. در فصل دوم مدل مارکوف پنهان و سایر مدلهای رایج را شرح می‌دهیم.

۱-۵ رویکرد این پروژه در جهت اعمال بهبود در بازشناسی گفتار

کیفیت و کمیت دادگان گفتاری تاثیر بسزایی در بهبود سیستم و الگوریتم های کاربردی بازشناسی گفتار پیوسته دارد. تا کنون تحقیقات وسیعی به منظور تهیه دادگان در زبان‌های مختلف انجام گرفته است. اکثر دادگان‌های رایج در این زبانها مبتنی بر واج هستند. فارس دات، مطرح ترین دادگان برای گفتار پیوسته در زبان فارسی نیز بر مبنای آموزش واج‌های فارسی طراحی شده است [۳].

همانگونه که در بخش ۱-۲ شرح داده شد واج‌ها به عنوان واحد گفتاری در سیستم های ناوابسته به متن به کار می‌روند. اما در سیستم های وابسته به متن که اثر واج‌های متوالی بر روی یکدیگر را در نظر می‌گیرند، از دو آوایی‌ها که اثر آوای سمت راست و یا اثر آوای سمت چپ را نیز لحاظ می‌کنند و یا از سه‌آوایی ها که اثرات آواهای سمت چپ و سمت راست را نیز در مدلسازی دخیل می‌کنند استفاده می‌شود. هجا و نیم‌هجا واحدهای گفتاری وابسته به متن می‌باشند. زبان فارسی دارای ساختار هجایی ساده ای می‌باشد و طبق تعریف برخی زبانشناسان، دارای سه نوع هجا می‌باشد [۳]:

- ساختار همخوان- واکه(CV)
- ساختار همخوان- واکه- همخوان(CVC)
- ساختار همخوان- واکه- همخوان- همخوان(CVCC)

هجاها دارای اطلاعات خوبی از لحن، آهنگ و نرخ گفتار می‌باشند. در زبان فارسی حدود ۶۰۰۰ هجا وجود دارد. مشکل اصلی استفاده از دو آوایی ها، سه آوایی ها و هجاها، تعداد زیاد این واحد ها در زبان فارسی می‌باشد که موجب پیچیده و زمان‌بر شدن فرآیند جستجو می‌شود. مشکل دیگر استفاده از این واحدها، کمبود داده های آموزشی به دلیل تنوع آنها و نیاز به داده های آموزشی زیاد می‌باشد.

[1] Hidden Markov Modeling

هر هجا به دو نیم‌هجای ابتدایی و انتهایی تقسیم می‌شود. نیم‌هجای ابتدایی شامل همخوان ابتدایی و بخشی از هسته هجا و نیم‌هجای انتهایی شامل باقیمانده هسته تا انتهای هجا می‌باشد. تعداد نیم‌هجاهای موجود در زبان فارسی که به صورت CV ، VC و VCC می‌باشند حدود ۷۰۰ نیم‌هجا می‌باشد. با توجه به پوشش مناسب و تعداد حالتهای کم نیم‌هجاها، استفاه از این واحدهای آکوستیکی کارآیی سیستم های بازشناسی گفتار پیوسته را بالا می برد.

یکی از اهداف این پروژه در راستای بهبود کارآیی سیستم بازشناسی گفتار پیوسته فارسی، تهیه متن یک پایگاه داده مناسب بر پایه نیم‌هجاهای رایج در زبان فارسی می‌باشد. شرح دقیق ملاحظات به کاررفته در تهیه متن این پایگاه داده در فصل سوم ارائه خواهد شد. در این پروژه واحد نیم‌هجا به عنوان واحد گفتاری در نظر گرفته شده است. یکی دیگر از مزایای این واحد گفتاری نسبت به سایر واحدهای آکوستیکی به ویژگی های واکه ها برمی‌گردد. بررسی منحنی انرژی سیگنال گفتار نشان می‌دهد، واکه ها در اکثر موارد از نظر فیزیکی به خوبی تلفظ می‌شوند و اثر آکوستیکی خوبی از خود باقی می‌گذارند. همچنین از منحنی انرژی سیگنال صحبت مشاهده می‌شود که انرژی متوسط واکه ها ، ۱۰dB تا ۱۵dB بیشتر از انرژی متوسط سیگنال در سایر بخش های گفتار می‌باشد. ویژگی دیگر واکه ها پریودیک بودن آنها برای تمامی گویندگان می‌باشد. این ویژگی ها سبب می‌شوند که تشخیص واکه ها و تخمین مرز آنها نسبت به همخوان ها آسان‌تر و دقیق‌تر صورت گیرد. لذا این واحد آکوستیکی همزمان از مزایای پوشش مناسب، تعداد حالتهای کم نیم‌هجاها در زبان فارسی و پایداری نسبی واکه نسبت به شرایط محیطی و سادگی آشکارسازی آن بهره‌مند می‌شود.

بنابراین برای ساده سازی بازشناسی گفتار با استفاده از مدل آماری مارکوف پنهان، ابتدا با استفاده از سکوت بین کلمات در یک جمله، مرزهای کلمات را مشخص می‌نماییم. آنگاه با بهره‌مندی از ویژگی های واکه‌ها، با تعیین مرز تقریبی آنها در هر کلمه و سپس طبقه بندی آنها توسط مدل آماری، جمله را به زیر بخش های واحد آکوستیکی تقسیم می‌کنیم. در آخر هر یک از زیر بخش‌ها را به کمک مدل آماری مارکوف پنهان بازشناسی می-کنیم. در فصل سوم به بررسی ویژگی‌های واکه‌ها که موجب سهولت و دقت در تشخیص و قطعه بندی واکه ها و نهایتاً منجر به انتخاب واحد آکوستیکی نیم‌هجا شده‌اند، می‌پردازیم.

۱-۶ ساختار پایان نامه

در فصل ۲ مدل‌های آماری رایج برای بازشناسی گفتار مورد بررسی قرار می‌گیرند. از آنجاییکه مدل آماری استفاده شده در این پروژه مدل مارکوف پنهان می‌باشد، پارامترهای این مدل و الگوریتم های رایج جهت پیاده سازی آن با جزئیات بیشتری نسبت به سایر مدلها شرح داده می‌شود. در ادامه نکات قوت و ضعف این مدل در بازشناسی گفتار و مدلهایی که جهت رفع این کاستی ارائه شده اند معرفی می‌شوند.

در فصل ۳ ابتدا به بررسی واحد های زبانی زبان فارسی و مزایا و مشکلات استفاده از هر یک از آنها برای کاربرد بازشناسی گفتار می‌پردازیم. سپس دلایل انتخاب واحد گفتاری نیم‌هجا برای بازشناسی گفتار پیوسته فارسی را مورد بررسی قرار می‌دهیم. در پایان نگاهی اجمالی خواهیم داشت بر ویژگیهای متن پایگاه دادگان تهیه شده در آزمایشگاه تحقیقاتی پردازش اطلاعات که با هدف پوشش کلیه نیم هجا های زبان فارسی تهیه شده است.

فصل ۴ به ارائه نتایج استفاده از انرژی میانگذر و مدل ترکیبی آکوستیکی آماری اختصاص دارد. در این فصل ابتدا به بررسی ویژگیهای واکه ها و مزایای استفاده از این ویژگیها برای بازشناسی واکه ها در گفتار پیوسته و گسسته می‌پردازیم. سپس نتایج استفاده از انرژی میانگذر بر روی گفتار گسسته و پیوسته را ملاحظه می‌کنید. در ادامه نتایج پیاده سازی دو روش ترکیبی آکوستیکی و آماری مارکوف پنهان بر روی دادگان گفتاری گسسته و پیوسته ارائه شده است. در انتهای این فصل جمع بندی و پیشنهادات برای ادامه این پروژه نیز مطرح شده است.

۲-۱ پیش گفتار

همانطور که پیشتر گفته شده با افزایش پیچیدگی سیستم های بازشناسی گفتار، تهیه قالب برای هر یک از دادگان بانک اطلاعات کاری غیر عملی می‌باشد. لذا از روشهای ریاضی برای تهیه مدل ها استفاده می‌نماییم. در این فصل در ابتدا روش غیر آماری پیچش زمانی پویا[1] و سپس روشهای آماری رایج در بازشناسی گفتار مطرح می‌شوند. با توجه به موفقیت مدل آماری مارکوف پنهان در بازشناسی گفتار، این مدل رایج ترین مدل آماری مورد استفاده برای بازشناسی گفتار و روش به کار رفته در این پروژه می‌باشد. بنابراین این مدل را با جزئیات بیشتری نسبت به سایر مدل ها شرح می‌دهیم.

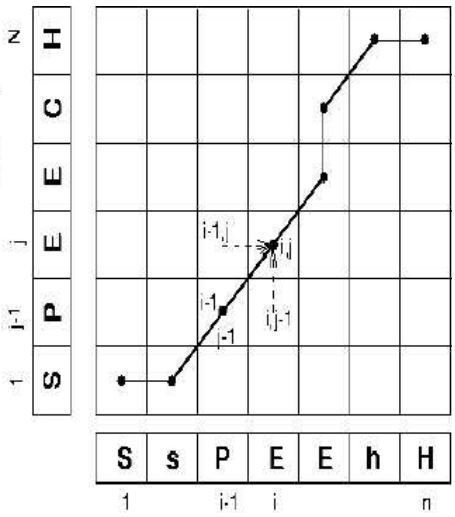
۲-۲ پیچش زمانی پویا

این روش یکی از اولین روش های مورد استفاده در بازشناسی گفتار و تنها مدل غیر آماری در این زمینه به شمار می‌رود[۱]و[۴]. پیچش زمانی پویا یک روش برنامه ریزی پویا می باشد که از مفهوم تطبیق الگو استفاده می‌کند. طول زمانی بیان های مختلف یک کلمه توسط افراد مختلف و حتی توسط یک فرد در بیان های گوناگون یکسان نیست. چنانچه گوینده ای کلمه ای را چند بار تکرار نماید، طول هر یک از واجهای ادا شده درکلمه مورد نظر با طول همان واج در تکرار دیگر از این کلمه متفاوت خواهند بود و ممکن است در برخی موارد گوینده یک واج را نسبت به حالت قبل کشیده تر و یا فشرده تر ادا کند. به همین دلیل اگر تعداد فریم های گویش اول M و

[1] Dynamic Time Warping (DTW)

تعداد فریم های گویش دوم N باشد، نمی توان فریم i ام از گویش اولی را با یک تابع خطی با فریم j ام از گویش دوم متناظر کرد. بنابراین برای استفاده از این روش در بازشناسی گفتار، ابتدا باید گفتار آزمایشی از نظر زمانی نسبت به گفتار مرجع تنظیم شود.

لذا باید از یک تابع غیر خطی برای تطبیق الگو استفاده نمود. الگوریتم معروف دایکسترا[1] برای یافتن کوتاه ترین مسیر بین دو گره یک گراف و تطبیق الگو ها به کار می‌رود. شکل (۲-۱) نمونه ای از عملکرد این روش را برای منطبق کردن قسمتهای مختلف از دو گویش یک کلمه را نشان می دهد.



شکل (۲-۱). تطبیق دو بیان از یک کلمه با استفاده از DTW

این روش تعیین می کند که بردار ویژگی i_x از گویش اول کلمه به طور بهینه به کدام بردار ویژگی از گویش دوم این کلمه شبیه است و اختلاف کمتری با آن نسبت به سایر بردارهای ویژگی گویش دوم دارد. به عبارتی یک تابع پیچش زمانی با استفاده از تکنیک برنامه ریزی پویا، در فضای دو بعدی اندیس i_x ازگویش اول را با اندیس i_y از گویش دوم متناظر می کند. در واقع روش DTW برای تطبیق زمانی دو کلمه و به دست آوردن میزان اختلاف آنها به کار می رود. از این روش می توان برای تطبیق بردارهای ویژگی مجموعه آزمون با بردارهای ویژگیهای مجموعه آموزش که به عنوان مرجع در نظر گرفته می‌شوند، استفاده نمود. یکی از نقاط ضعف DTW نیاز به داشتن مدل های مرجع برای مقایسه می‌باشد. با افزایش دامنه لغات از مدلهای زیر لغوی مانند آوا به جای

[1] Dijkstra

کلمات استفاده می‌شود. اما تهیه مدل های مرجع مناسب از آوا ها با توجه تغییرات آنها وابسته به متن کاری مشکل می‌باشد. عیب اصلی این روش، کند بودن آن است. اگرچه روش‌های تکاملی برای افزایش سرعت DTW مورد استفاده واقع شده است، اما در هنگامی که تنوع گفتارهای ورودی یا تعداد واحدهای زبانی مورد بازشناسی زیاد باشد، استفاده از DTW در بازشناسی گفتار مناسب نیست.

۲-۳ شبکه های عصبی[1]

شبکه های عصبی با توجه به توان بالا در پردازش موازی و محلی ، قابلیت یادگیری، تعمیم، طبقه بندی، به خاطر سپردن و به خاطر آوردن الگوها به صورت محتوایی خیزش وسیعی را در زمینه هایی از قبیل سیستم های هوشمند مصنوعی، شناسایی و کنترل تطبیقی سیستم ها، سیستم های شناسایی گفتار، تصویر و ... ایجاد کرده است[۸- ۴].

شبکه های عصبی به جای انجام مراحل الگوریتمی و قدم به قدم، قوانین انجام کار را برای خود تولید می-نمایند. به بیان دیگر شبکه با روش سعی و خطا، به تدریج نحوه انجام کار را یاد می‌گیرد. عملکرد شبکه های عصبی بر اساس شبیه سازی عملکرد مغز انسان می‌باشد. شبکه عصبی مجموعه‌ای است از عناصر ساده محاسباتی که دارای اتصالات زیادی با یکدیگر می‌باشند. هر عنصر مجموعه که دارای چندین ورودی و یک خروجی است را " نرون" و اتصالات ما بین آنها را " سیناپس" می نامند. فرض کنید که بردار ویژگیهای مورد نظر یا ورودی ها بصورت $x = [x_1, x_2, ..., x_N]^T$ باشد، آنگاه در هر نرون که هسته اصلی یک شبکه عصبی می‌باشد، ورودی x و خروجی y با رابطه (۲-۱) به هم مربوط می‌باشند:

$$y = f\left(\sum_{i=1}^{N} w_i x_i - Q\right) \qquad (۱-۲)$$

یک نرون در واقع یک ابر صفحه در فضای ویژگی است که دو ناحیه را با یک صفحه از هم جدا می کند. مقدار Q مقدار سطح آستانه نرون است و W_i ها بردار وزن هستند و بردار y صفحه را تشکیل می‌دهد. تابع f، تابع

[1] Neural Networks

فعالیت واحد عصبی است که به عنوان مثال می‌تواند یک تابع پله، تانژانت هیپربولیک و یا یک تابع سیگموئید باشد. چند نمونه تابع فعالیت که دارای مقادیر بین صفر و یک هستند، در زیر نشان داده شده اند:

$$f(x)=\begin{cases}+1 & x\geq 0\\ -1 & x<0\end{cases} \quad , \qquad f(x)=tgh(\beta x) \qquad \beta>0 \quad , \quad f(x)=\frac{1}{1+e^{-x}}$$

در فضای ویژگی از ابرصفحه ها برای نمایش مرزهای یک ناحیه با اطرافش استفاده می‌کنند و در واقع با تعدادی از واحدهای عصبی ناحیه‌ای از فضای بردار ویژگی را از سایر نواحی جدا می‌نمایند. چنانچه هر کلاس از کلاسهای ویژگی را با تعدادی نرون مدلسازی نماییم، عمل دسته بندی یا جدا کردن یک دسته از دسته‌های دیگر با تعداد زیادی از واحدهای عصبی صورت می‌گیرد.

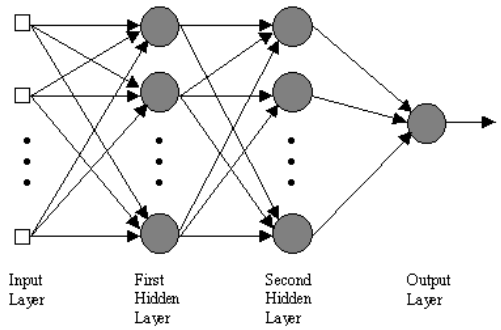
یکی از ویژگی های عمده شبکه عصبی این است که با استفاده از آن می‌توان تابع پیچیده غیر خطی را پیاده سازی نمود و فضا را به نواحی مختلف تقسیم نمود. در این شبکه ها نیازی به شناخت الگوریتم تولید خروجی از روی داده ورودی وجود ندارد و چنانچه داده ورودی اختلاف اندکی با ورودی مناسب برای تولید یک خروجی مشخص داشته باشد، سیستم اختلاف را در نظر نگرفته و خروجی مطلوب را ایجاد می‌کند. در حقیقت سیستم یک نگاشت بین داده های ورودی و خروجی ایجاد می‌کند.

شبکه های عصبی را می‌توان به دو دسته پس خورد[1] و پیش خورد[2] تقسیم نمود. اگر خروجی نرونهای یک لایه به ورودی نرونهای لایه قبل متصل باشد شبکه عصبی از نوع پیش خورد خواهد بود. برای استفاده از شبکه های عصبی ابتدا باید آنها را آموزش داد. در فرآیند آموزش، با استفاده از یک سری داده آموزشی، وزنهای اتصالات نرونها تنظیم می‌شوند. در فرآیند آموزش اگر دادهای آموزشی شامل زوجهای ورودی خروجی باشد شبکه از نوع یادگیری با سرپرستی می‌باشد. یکی از شبکه های عصبی معروف در پردازش گفتار شبکه عصبی پرسپترون[3] نام دارد.

[1] feedback
[2] feed forward
[3] perceptron

- **شبکه پرسپترون چند لایه[1] (MLP)**

شبکه MLP یک شبکه متشکل از لایه های ورودی ، مخفی و خروجی است که به ترتیب قرار گرفتن واحدهای عصبی در آن در شکل (۲-۲) نشان داده شده است. گره های ورودی این شبکه به هر یک از المانهای بردار ویژگی متصل خواهند شد. گره های لایه های مخفی هر کدام بیانگر یک ابر صفحه می‌باشند و هر یک از گره های خروجی متناظر با یکی از کلاسهای فضای ویژگی هستند که در صورت تعلق بردار ویژگی به آن دسته، گره نظیر آن فعال می‌گردد. البته ممکن است شبکه MLP دارای چند لایه مخفی نیز باشد.



شکل (۲-۲).شبکه پرسپترون چند لایه

الگوریتم یادگیری و تصحیح وزنها در شبکه عصبی MLP، الگوریتم انتشار خطا به عقب[2] می‌باشد که همواره در صدد مینیمم کردن خطاهای دسته بندی بردارهای ویژگی است. شبکه MLP استاندارد در حالت عادی یک شبکه استاتیک است و دینامیک بودن و تغییرات زمانی گفتار را در نظر نمی‌گیرد و این شاید تنها ایرادی باشد که بتوان بر آن گرفت. ولی یکی از مزایای شبکه MLP قدرت تمایز بین کلاسها است.

- **شبکه عصبی با تاخیر زمانی (TDNN)**

همانطور که قبلا اشاره شد سیگنال گفتار طبیعتی غیر ایستان دارد و مشکل اکثر روش ها آن است که طبیعت دینامیک سیگنال را در نظر نمی‌گیرند. شبکه عصبی با تاخیر زمانی راهی برای در نظر گرفتن

[1] Multi Layer Perceptron
[2] Back propagation