فهرست مطالب

عنوان **صفحه**

فهرست شکل‌ها

عنوان **صفحه**

فصل ۱ - مقدمه

۱-۱- پیشگفتار

تکنیک استفاده از ویژگیهای منحصر به فرد زیستی انسانها برای بازشناسی آنان دیر زمانیست که در حیطه‌ی جرم‌شناسی به عنوان ابزاری مطمئن و کارآمد مطرح می‌باشد، به لحاظ باور علمی این مطلب که این ویژگیها غیر قابل تقلید بوده، احتمال مشابهت آنها در افراد صفر یا عددی مشابه آن است. متخصصان طراحی سیستمهای امنیتی الکترونیکی نیز در دهه‌های اخیر به این فن‌آوری به‌عنوان مهمترین اساس برای طراحی سیستمهای امنیتی وابسته به فرد خاص نظر داشته‌اند.

طی این سال‌ها تلاش زیادی روی بازشناسی‌صحبت صورت گرفت. اما با توجه به عوامل زیادی که در آن موثر هستند، همواره عملیات تشخیص با خطا روبه‌رو بوده است. تارهای صوتی انسان خصوصیاتی غیرخطی دارند و از طرف دیگر عملیات آنها کاملا در اختیار نیست، بلکه عوامل مختلفی از جنسیت تا حالت عاطفی فرد در آن تاثیرگذار است. در نتیجه تلفظ صوتی می‌تواند به لهجه، طرز تلفظ، طرز گفتار و میزان شمرده بودن آن، درشتی صدا، تو دماغی حرف زدن، زیر و بمی صدا، بلندی صدا و سرعت ادای کلمات بستگی داشته باشد. علاوه بر اینها از آنجا که معمولا افراد در محیطی صحبت می‌کنند که صداهای محیطی نیز وجود دارد، این مسئله پیچیده‌تر می‌شود به شکلی که بازشناسی‌صحبت حتی از تولیدصحبت سخت‌تر و پیچیده‌تر می‌شود.

۱-۲- شیوه‌های نوین

گرچه بسیاری از سیستم‌های بازشناسی‌صحبت در محیط‌های‌بدون نویز به نتایج رضایت بخشی رسیده اند. ولی بااین‌حال یکی از بزرگترین مسائل در حوزه‌ی‌بازشناسی‌صحبت مسئله‌ی‌دقت بازشناسی است. زیرا اگر محیط‌آموزش از محیط‌آزمایش متفاوت باشد دقت کم خواهد شد. این اختلافات محیطی به دلایلی همچون نویزجمعی، انحراف‌کانال، اختلافات‌صوتی میان گوینده‌های‌مختلف و غیره می‌باشد.

الگوریتم‌های زیادی برای کمک به مقاوم‌سازی محیط سیستم‌های بازشناسی‌صحبت مطرح شده‌اند. ساده‌ترین فرم نرمالیزه کردن محیط، نرمالیزاسیون‌متوسط‌کپسترال[1] (CMN) می‌باشد[5،4]، که باعث می‌شود متوسط‌کپسترال برای تمامی گفته‌ها به سمت صفر میل کند. همچنین اگر پاسخ ضربه در مقایسه با طول فریم آنالیز شده کوتاه باشد، موجب حذف فیلترینگ خطی ایستان خواهد شد. نرمالیزاسیون

[1] Cepstral Mean Normalization

واریانس‌متوسط[1] (MVN) را می‌توان به‌عنوان بسطی از CMN در نظر گرفت[6،5]. در MVN متوسط و واریانس بردارهای‌کپسترال بایستی برای تمامی‌گفته‌ها به ترتیب به صفر و یک نرمالیزه شود. در بیشتر نمونه‌های عملی از نرمالیزاسیون هیستوگرام[2] استفاده می‌شود. در این روش فرض می‌شود بردارهای کپسترال همه‌ی گفته‌ها، تابع چگالی احتمال یکسان داشته باشند. اخیرا، محققان به این نتیجه رسیده‌اند که نرمالیزاسیون هیستوگرام روی دلتا-کپسترال و نیز ضرایب کپسترال اصلی باعث بهبودهای بیشتری در عملکرد بازشناسی می‌شود[7].

دومین گروه از روشها مبتنی بر تخمین مؤلفه‌های نویز برای کلاسترهای مختلف و سپس استفاده از این اطلاعات برای تخمین طیف صحبت اصلی می‌باشد. نرمالیزاسیون کپسترال وابسته به کلمه‌ی کد[3] (CDCN) [8] و سری‌های تیلور برداری[4] (VTS) [9] مثال‌هایی از این گروه می‌باشند. این الگوریتم‌ها ممکن است برای کاهش طیفی در نظر گرفته شوند[10]، و باعث کاهش طیفی در بعد کپسترال شوند.

ولو اینکه برخی از این الگوریتم‌ها، بهبودهایی را برای نویز ایستان نشان می‌دهند[12،11]. با این حال، بهبود در نویز غیرایستان همچنان به قوت خود باقیست[13]. اخیرا روشهایی مبتنی بر پردازش صوت انسان[14] و نیز روشهایی مبتنی بر ویژگی‌های گم شده[5][15]، مطرح شده‌اند که می‌تواند در حل این مسئله نویدبخش باشد. در [14] دقت بازشناسی صحبت با استفاده از یک مدل مطمئن از پردازش صوتی انسان در عصب شنیداری، بهبود میابد.

سومین روش، جداسازی سیگنال مبتنی بر آنالیز اختلاف در زمان حضور می‌باشد[18،۱۷،۱۶]. به خوبی پیداست که سیستم دو گوشی بودن[6] انسان در قابلیت جداسازی صحبت رسیده از زوایای مختلف رسیدن به گوشها قابل توجه است[18]. مدل‌های زیادی برای توصیف دو گوشی بودن انسان مطرح شده اند[20،۱۹]، که ازجمله می‌توان به اختلاف زمان رسیدن به گوش[7] (ITD)، اختلاف فاز رسیدن به گوش[8] (IPD)، اختلاف شدت رسیدن به گوش[9] (IID)، و تغییرات کرولیشن رسیدن به گوش[10] اشاره کرد. جدیداً، الگوریتم تخمین دامنه‌ی عبور از صفر[11] (ZCAE) توسط park مطرح شده است[17]. این

[1] Mean Variance Normalization
[2] histogram normalization
[3] Codeword Dependent Cepstral Normalization
[4] Vector Taylor Series
[5] missing-feature
[6] binaural system
[7] interaural time difference
[8] interaural phase difference
[9] interaural intensity difference
[10] Changes of interaural correlation
[11] Zero Crossing Amplitude Estimation

الگوریتم‌ها (و البته الگوریتم‌های مشابه دیگر) نوعا، صحبت وارد شده در کانالهای میان‌باند را بررسی می‌کند و تلاش دارد زیرمجموعه مؤلفه‌های فرکانسی-زمانی را شناسایی کند، برای آنکه ITD نزدیک به ITD نامی منبع صدای موردنظر باشد.

۱-۳- هدف از انجام تحقیق

بسیاری از الگوریتم‌های نرمالیزاسیون ساده در بعدویژگی (کپسترال) یا زمان عملکرد رضایت بخشی داشته‌اند، بااین‌حال استفاده‌از نرمالیزاسیون در بعدطیفی یا توان فواید خاص خود را دارد: اول اینکه، نرمالیزاسیون‌طیفی می‌تواند به‌راحتی به‌عنوان یک طبقه‌ی‌پیش‌پردازش برای بسیاری از سیستم‌های استخراج ویژگی استفاده شود و دوم اینکه، این نوع نرمالیزاسیون می‌تواند به‌عنوان قسمتی‌از یک طرح بهبود صحبت استفاده شود[2،1]، ما در این پایان‌نامه از نرمالیزاسیون در بعدطیفی استفاده کردیم.

و در نهایت می‌توانیم بگوییم که هدف از این پایان‌نامه، شرح الگوریتم مقاوم به نویزی است که از سیستم شنیداری انسان ناشی شده است و به کمک این الگوریتم می‌توان بازشناسی صحبت را بطور چشمگیری افزایش داد.

۱-۴- نوآوری تحقیق

در آنالیز زمان-فرکانس، مقدار طول پنجره‌ی بهینه را برای جبران نویز مطرح می‌کنیم. همچنین مزایای پنهانی را که با اختصاص دادن وزن فرکانس بدست میآید، بحث می‌کنیم. و همچنین یک روش کارا از نرمالیزه کردن مؤلفه‌های نویز مبتنی بر این مشاهدات را معرفی می‌کنیم. درحالی که، ارتباط شدت یک صدا بر بلندی درک شده‌ی آن معلوم است، تلاشهای زیادی برای تحلیل تأثیرات نسبت سطح غیرخطی انجام شده است. در این پایان‌نامه، چندین غیرخطی را که از عصب شنیداری انسان ناشی می‌شود، معرفی می‌کنیم و نشان می‌دهیم که غیرخطی تابع توان به مراتب قدرتمندتر از غیرخطی لگاریتمی -که در اکثر سیستم‌های استخراج ویژگی صحبت(MFCC) استفاده شده است- می‌باشد[21].

موضوع مهم دیگری که در کارمان از آن استفاده می‌کنیم، نرمالیزه کردن توان است. یکی‌از مشخصات سیگنالهای صحبت این است که سطح توانشان بسیار سریع تغییر می‌کند درحالی که تغییرات نویز پس زمینه معمولا بسیار آهسته می‌باشد. در نویز ایستان همانند نویز صورتی یا سفید، تغییرات توان به صفر می‌رسد اگر طول پنجره‌ی‌آنالیز بطور قابل‌توجهی بزرگ باشد. حتی در نمونه‌ی نویز غیرایستان همچون نویز موسیقی، توان نویز به سرعت توان صحبت تغییر نمی‌کند. به‌همین‌دلیل برای تشخیص اینکه فریم جاری نویزی شده است یا خیر می‌توان از توزیع توان استفاده کرد. یک روش موثر برای انجام اینکار،

اندازه‌گیری نسبت میانگین‌حسابی به میانگین‌هندسی[1](نسبت AM-GM) است[22]. زیرا اگر مقادیر توان خیلی سریع تغییر نکند، میانگین‌حسابی و هندسی مقادیری شبیه خواهند داشت. ولی اگر تغییرات توان سریع باشد، میانگین حسابی بسیار بزرگتر از میانگین هندسی خواهد بود.

در این پایان‌نامه نرمالیزاسیون توزیع‌توان باطول‌متوسط را معرفی می‌کنیم که باعث دقت بیشتر در بازشناسی‌صحبت می‌شود. به‌این دلیل که سیستم‌های ادراکی روی تغییرات سیگنال هدف متمرکز شده اند و سطوح پس‌زمینه‌ی ثابت را نادیده می‌گیرند، الگوریتمی که در این پایان‌نامه معرفی می‌کنیم از بعضی جهات شبیه کاهش‌طیفی می‌باشد منتها به‌جای تخمین توان‌نویز از قسمت‌های غیرصحبت یک گفته، بایاسی را کم می‌کنیم که برای نمایش یک سطح ناشناخته از شبیه‌سازی پس‌زمینه فرض‌شده‌بود. یکی از روش‌های انجام اینکار، کاهش‌بایاس‌توان باطول‌متوسط مبتنی‌بر نسبت میانگین‌حسابی به میانگین هندسی (نسبت AM-GM) است[22]. یعنی اگر از روشهای آنالیز موقتی همچون روش متوسط‌جاری و روش پنجره‌باطول‌متوسط استفاده کنیم، می‌توانیم نویز را به‌میزان مناسبی جبران کنیم. در این پایان‌نامه، روش دیگر کاهش‌بایاس‌توان[2] (PBS) را بااستفاده‌از روش متوسط‌جاری باطول‌متوسط[3]، پیشنهاد می‌دهیم.

در این پایان‌نامه، یک الگوریتم نرمالیزاسیون جدید را که مبتنی بر اصل نسبت AM-GM می‌باشد معرفی کرده و با ماکزیمم کردن تیزی توزیع توان دقت سیستم بازشناسی‌صحبت را به کمک آن بهبود می‌دهیم. همانطور که گفتیم نسبت میانگین‌حسابی به میانگین‌هندسی توان در یک باند فرکانسی مخصوص بستگی به‌میزان نویز در محیط دارد. با استفاده از مقادیر AM-GM بدست آمده از پایگاه داده‌ی صحبت تمیز، می‌توان از یک تبدیل غیرخطی(بطور خاص یک تابع توان) برای تبدیل توان‌های خروجی استفاده کرد. زیرا نسبت AM-GM در هر باندفرکانس‌ورودی نسبت مشابهی را نتیجه می‌دهد. این نسبت را در صحبت تمیزی که برای آموزش سیستم نرمالیزاسیون استفاده می‌شود، مشاهده می‌کنیم.

۱-۵- ساختار گزارش

این پایان‌نامه به‌صورت زیر تنظیم شده است: در فصل 2- مفاهیم کلی در یک سیستم بازشناسی صحبت مطرح می‌شود. سپس در فصل 3- خلاصه‌ای کلی از سابقه‌ی تئوری‌های انجام شده‌ای که در راستای این پایان‌نامه است، مطرح می‌شود و بطور اجمالی از مفاهیم کلی و مؤثر از هر ایده و الگوریتمی که در این حوزه است بحث می‌شود. در فصل 4- تحلیل فرکانسی و زمانی و تأثیر آن در بازشناسی صحبت بحث می‌شود. خواهیم دید که طول پنجره و وزن فرکانسی اثر قابل توجهی در دقت بازشناسی

[1] arithmetic mean to geometric mean ratio

[2] Power Bias Subtraction

[3] medium-duration running average method

صحبت دارد. در ادامه‌ی این فصل غیرخطی بودن شنیداری و چگونگی تأثیر آن در سیستم‌های بازشناسی صحبت مقاوم به نویز را بررسی می‌کنیم. غیرخطی بودن شنیداری، ارتباط میان شدت صدا و نمایش چگونگی پردازش شنیداری است که نقش بسیار مهمی را در بازشناسی صحبت بازی می‌کند. در فصل 5- یک الگوریتم استخراج ویژگی جدید را که ضرایب کپسترال نرمالیزه شده‌ی توان[1] (PNCC) نامیده می‌شود، معرفی می‌کنیم و به کمک ماکزیمم کردن تیزی توزیع توان باعث بهبود بازشناسی می‌شویم. نهایتا در فصل 6- نتایج شبیه‌سازی را نشان می‌دهیم و در نهایت به بیان نتیجه‌ی کلی می‌پردازیم.

[1] Power Normalized Cepstral Coefficient

فصل ۲ - کلیات سیستم بازشناسی گفتار

۲-۱- مقدمه

سیگنال گفتار را می‌توان به‌صورت یک پوش طیفی که به‌آرامی تغییر می‌یابد، در نظر گرفت. این پوش طیفی به‌وسیله‌ی انسان دریافت می‌گردد و به دنباله‌ای از کلمات و معانی آنها ترجمه می‌گردد. سیستم‌های بازشناسی خودکار گفتار نیز بطور مشابه تلاش می‌کنند که این پوش طیفی را به دنباله‌ای از کلمات تبدیل نمایند. مشکلات فراوانی مانند تغییرات پوش طیفی گفتار در این راه وجود دارد. تغییرات پوش طیفی گفتار به دلایلی نظیر جنسیت گوینده، نحوه بیان و تأکید در گفتار گوینده و نیز تغییر محیط آکوستیکی که سیستم بازشناسی در آن عمل می‌کند، به‌وجود می‌آید. طراحی سیستم بازشناسی خودکاری که بتواند با توانایی انسان در مقابله با این تغییرات برابری نماید، هنوز یک مسأله و چالش اساسی محسوب می‌شود. بسیاری از سیستم‌های بازشناسی گفتار از روشهای آماری برای مقابله با دسته‌ای از تغییرات پوش طیفی استفاده می‌کنند. کارایی این سیستمهای آماری تا به امروز به‌طور قابل توجهی افزایش یافته است، به‌طوریکه برای یک کتاب لغت نامحدود مستقل از گوینده صحت بازشناسی بیش از 90% بدست آمده است. به‌علاوه، برای کتاب لغت‌هایی با اندازه‌ی محدود، صحت بازشناسی بیش از 95% نیز قابل دسترسی است. با این درصد کارایی، این سیستمها قابل بهره برداری به‌نظر می‌رسند، ولی باید توجه کرد که اکثر این سیستمها در محیط‌هایی یکسان و ساکت (بدون حضور نویز) مورد آموزش و آزمایش قرار گرفته‌اند. این در حالی است که در شرایط عملی، سکوت به‌ندرت وجود دارد و محیط‌های آکوستیکی آزمایش و آموزش نیز عموماً با یکدیگر متفاوتند.

هدف در بحث بازشناسی گفتار گسسته و پیوسته‌ی فارسی، دستیابی به سیستم‌های بازشناسی گفتار فارسی با توانایی بازشناسی گفتار گسسته در چارچوب دایره‌ی کلمات تعیین شده و نرخ بازشناسی قابل قبول، بر روی دادگان تعریف شده و نیز بازشناسی گفتار پیوسته‌ی فارسی با دایره‌ی کلمات مشخص و در چارچوب دادگان اصلاح شده‌ی پیوسته و با نرخ بازشناسی مورد نظر می‌باشد. در هر یک از این دو راستا، اقدامات صورت گرفته شامل مطالعه، بررسی و پیاده‌سازی روش‌های مطرح در بازشناسی جهت دستیابی به نتایج مطروحه بوده است.

در یک سیستم بازشناسی گفتار، پس از تقسیم سیگنالها و نمونه‌های گفتاری به دو دسته‌ی دادگان آموزش و آزمایش، سیگنال ورودی به فواصل زمانی 25 تا 30 میلی‌ثانیه (با در نظر گرفتن درصدی همپوشانی) قاب‌بندی می‌شود. سپس در مرحله‌ی استخراج ویژگی، از هر قاب ویژگی‌های گفتاری (ضرایب

تخمین خطی[1] و ضرایب کپسترال مبتنی بر معیار مل[2]) استخراج می‌گردند. در مرحله‌ی آموزش، از این ویژگیها برای آموزش مدل‌های آماری بازشناسی نظیر مدل مخفی مارکف[3] (HMM) و شبکه عصبی(ANN) استفاده می‌شود. در مرحله‌ی آزمایش یا بازشناسی، ویژگی‌ها از طریق یک الگوریتم مصنوعی بازشناسی نظیر ویتربی[4] با این مدل آماری مقایسه می‌شوند.

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

۲-۲- تاریخچه تحقیقات انجام شده در زمینه‌ی بازشناسی گفتار

اولین پژوهشهای صورت گرفته به منظور بازشناسی گفتار توسط ماشین و بطور اتوماتیک به بیش از 5 دهه قبل بر می‌گردد. این روند با تلاش برخی دانشمندان برای شناسایی مدلهای آکوستیکی آواها دهه ی1950 آغاز شد. در سال 1952 در آزمایشگاه بل، دیویس، بیدالف و Balashek سیستمی برای شناسایی ارقام گسسته ساختند که تنها بر پایه رزونانس حاصل از قسمتهای صدادار هر کدام از ارقام کار می‌کرد. در سال 1956 در طول یک تحقیق مستقل دیگر ، اوسلن و بلر توانستند در آزمایشگاه RCA ده هجای مختلف از صدای یک فرد خاص را بازشناسی کنند که در واقع کلمات مختلف یک هجایی بودند. این سیستم هم با معیارهای طیفی حاصل از بخشهای صدادار سیگنال گفتار کار می‌کرد. در سال 1959 در انگلستان ، فری و دنیس توانستند 5 حرف صدادار و 9 حرف بیصدا را با استفاده از یک آنالیز کننده طیف سیگنال گفتار و یک سیستم تطبیق الگو بازشناسی کنند. ایده‌ی جدیدی که در این سیستم مطرح شده بود عبارت بود از بهبود نتیجه بازشناسی با توجه به اطلاعات آماری واج‌هایی که می‌توانند کنار هم قرار بگیرند. تلاش قابل توجه دیگری که در این دوره انجام شد پیاده‌سازی یک سیستم بازشناسی حروف صدادار بود که در سال 1959 در آزمایشگاه لینکلن MIT انجام شد و Forgie and Forgie نام گرفت. در

[1] Linear Predictive Coding
[2] Mel Frequency Cepstral Coefficients
[3] Hidden Markov Model
[4] viterbi

این سیستم ده حرف صدادار به شرطی که در کلمه‌ای بین دو واج /ب/ و /ت/ قرار میگرفتند (طبق قالب /b/-vowel-/t) به‌صورت مستقل از گوینده تشخیص داده می‌شدند.

در دهه‌ی 1960 چندین ایده پایه‌ای در زمینه پردازش گفتار مطرح و منتشر شد. شروع این دهه همراه بود با ورود بسیاری از محققین آزمایشگاههای ژاپنی به عرصه پردازش گفتار و ارائه سیستم‌هایی که در آنها حتی سخت افزار مورد استفاده هم به‌صورت خاص و برای همین منظور طراحی شده بود. به‌عنوان یکی از نمونه‌های اولیه این سیستمها می‌توان به سیستم تشخیص حروف صدادار اشاره کرد که توسط سوزوکی و ناکاتا در آزمایشگاه تحقیقات رادیویی شهر توکیو اجرا شد. در این سیستم یک بانک فیلتر ، اطلاعات طیفی سیگنال صدا را استخراج کرده و خروجی هر فیلتر را از طریق یک کانال مجزا و به‌صورت وزندار به یک مدار منطقی تصمیم‌گیر می‌داد. تابع تصمیم بکار رفته یک تابع اکثریت بود که منجر به بازشناسی واج صدادار موردنظر می‌شد. نمونه‌ی دیگر، سیستم بازشناسی واجی بود که در سال 1962 توسط ساکایی و دوشیتا در دانشگاه کیوتو طراحی شد و با استفاده از یک سخت‌افزار بخش‌بندی کننده سیگنال گفتار به‌همراه آنالیز نرخ عبور از صفر برای بخشهای مختلف گفتار خروجی را تولید می‌کرد. یکی دیگر از تلاشهای انجام شده از سوی ژاپنی‌ها طراحی یک سیستم سخت افزاری برای بازشناسی ارقام بود که در سال 1963 در آزمایشگاه NEC[1] توسط ناگاتا و همکارانش صورت گرفت. این شاید اولین و قابل توجه‌ترین پژوهش انجام شده در این آزمایشگاه بود که بعدها به یک پروسه تحقیقاتی طولانی مدت و تولیدی تبدیل شد.

در دهه‌ی 1960 سه فعالیت اساسی در زمینه‌ی پردازش گفتار انجام شد که در واقع حاصل تمامی تحقیقات صورت گرفته در طول 20 سال قبل از آن بود. اولین مورد از این فعالیتها پروژه‌ای بود که به دست مارتین و همکارانش در آزمایشگاه RCA انجام شد و منجر به یافتن راه حلی واقعی برای مواجه با مشکلات ناشی از ناهمزمانی‌ها و غیر یکنواختی فرآیندهای گفتاری شد. در همین زمان فردی بنام وینتسیاک استفاده از روشهای پویا را برای انطباق بهینه‌ی دو دنباله‌ی گفتاری مطرح ساخت. اگرچه ایده‌های مربوط به الگوریتم‌هایی نظیرDTW[2] و روشهای بازشناسی کلمات پیوسته در نظریات وینتسیاک موجود بود اما این حقایق تا اوایل دهه‌ی 1980 به‌صورت عملی بروز نکرد. در این زمان روشهای ساده‌تری برای حل این مسائل توسط سایر افراد ارائه شده بود.

آخرین دستاورد مهمی که در اواخر دهه‌ی 60 ، تحقیقات منحصر بفرد ردی در زمینه‌ی بازشناسی گفتار پیوسته با استفاده از روشهای پویای جستجوی واج‌ها بود. این تحقیقات منجر به بنیانگذاری یک

[1] Nippon Electric Corporation

[2] Dynamic Time Warping

پروسه تحقیقاتی بسیار موفق در دانشگاه CMU[1] شد بطوریکه هنوز هم این دانشگاه از سردمداران طراحی سیستمهای بازشناسی گفتار پیوسته در دنیا می‌باشد.

در دهه‌ی 70 نیز چند نتیجه‌ی قابل توجه دست آمد. اول اینکه بحث بازشناسی گفتار گسسته با مطالعات ولیچکو و زاگورایکو در روسیه ، ساکو و چیبا در ژاپن و ایتاکورا در آمریکا مورد توجه و استفاده قرار گرفته و به‌عنوان یک تکنولوژی کاربردی مطرح شد. مطالعات روسها نشان داد که می‌توان از روش‌های بازشناسی الگو به‌خوبی در مسائل بازشناسی گفتار استفاده کرد ، تحقیقات ژاپنیها اثبات کرد استفاده از روشهای برنامه‌سازی پویا در اینگونه مسائل بسیار مفید و موفقیت آمیز خواهد بود و پژوهش‌های ایتاکورا نشان داد می‌توان ایده‌های کدگذاری پیشگوی خطی را ، که قبلا با موفقیت در کدگذاری گفتار با نرخ بیت پایین مورد استفاده قرار گرفته بود ، با استفاده از معیارهای فاصله‌ی مناسب در محدوده‌ی مسائل بازشناسی گفتار هم بکار برد. دومین پدیده‌ی قابل توجه دهه‌ی 70 ، شروع یک پروژه‌ی تیمی بسیار موفق و طولانی مدت در زمینه‌ی سیستمهای بازشناسی گفتار با لغتنامه‌های بزرگ در IBM بود.

سرانجام آزمایشاتی در زمینه‌ی سیستم‌هایی که کاملا مستقل از گوینده‌اند انجام شد که در طی آن بسیاری از روشهای خوشه‌بندی مورد استفاده قرار گرفت تا تخمینی برای تعداد الگوهای متفاوتی که لازم است تا سیستم بازشناسی گفتار موردنظر دامنه‌ی وسیعی از گوینده‌های مختلف را پوشش دهد بدست آید. این تحقیقات در طول 10 سال بقدری کامل و پخته شد که امروزه تکنیک‌های قابل استفاده برای تولید الگوهای مستقل از گوینده موجود و کاملا تحلیل شده‌اند.

در دهه‌ی 70 موضوع رایج در تحقیقات بحث بازشناسی گفتار گسسته بود ، در دهه‌ی 80 این روند به بازشناسی گفتار پیوسته ارتقا پیدا کرد. هدف در این زمان ایجاد سیستمی بود که بتواند در شرایط مختلف گفتار پیوسته و روان را با به‌دنبال هم آوردن مدلهای تک تک کلمات و با دقت کافی بازشناسی کند. در نتیجه الگوریتم‌های زیادی برای پیدا کردن کم هزینه‌ترین دنباله‌ی قابل ورودی قابل انطباق با الگوهای کلمات ارائه شد که هر کدام دارای ویژگی‌های پیاده‌سازی خاص خود بوده و در محدوده‌ی وسیعی از مسائل خاص خود کاربرد داشت.

در نتیجه تحقیقات دهه 80 ، تکنولوژی‌های برپایه‌ی الگو که قبلا بیشتر مطرح بود جای خود را به روشهای مدل‌سازی آماری و بخصوص مدلهای مخفی مارکوف داد. اگرچه در ابتدا این روش مدل‌سازی بجز در تعداد محدودی از آزمایشگاه‌ها ، در جایی از دنیا شناخته شده نبوده و اطلاعات مربوط به پیاده‌سازی آن در دست نبود ، اما با گذشت نیم‌دهه و انتشار مستندات مربوط به آن، چنان به سرعت

[1] Carnegie Mellon University

مورد استقبال قرار گرفت که در همه‌ی آزمایشگاه‌های سراسر دنیا از مدلهای مخفی مارکوف به‌صورت گسترده استفاده می‌شد. تکنولوژی دیگری که دوباره در دهه‌ی 80 مطرح شد ، استفاده از شبکه‌های عصبی در مسائل بازشناسی گفتار بود. البته ایده‌ی شبکه‌های عصبی از سالهای 1950 مطرح شده بود اما در شروع دارای مشکلات و کمبودهای عملی بود. اما با گذشت زمان قدرت این ابزار و توانایی آن در حل محدوده‌ی وسیعی از مسائل شناخته شد و پس از آن چندین سیستم با استفاده از این روش و به‌طرق مختلف پیاده‌سازی شده و مورد استفاده قرار گرفت.

در ادامه دهه‌ی 80 بازهم مباحث مربوط به لغت نامه‌های بزرگ مورد توجه خاص قرار گرفت و انجمن DARPA[1] مسئولیت مدیریت یک پایگاه داده‌ی دقیق و کامل از 1000 کلمه برای استفاده در کابردهای بازشناسی گفتار پیوسته را عهده‌دار شد. تحقیقات گسترده در دانشگاه CMU (که منجر به طراحی سیستم SPHINX شد)، آزمایشگاه لینکلن، MIT، SRI ، آزمایشگاه بل و بسیاری از مراکز تحقیقاتی دیگر شدت گرفت و تا دهه‌ی 90 ادامه پیدا کرد. همزمان با این فعالیتها تکنولوژی‌های بازشناسی گفتار به سرعت در شبکه‌های تلفنی برای اتوماسیون و بهبود خدمات اطلاع رسانی مورد استفاده قرار گرفت.

۲-۳- پارامترهای بازشناسی گفتار

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

۲-۳-۱- وابسته یا مستقل از گوینده

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

[1] Defense Advanced Research Projects Agency

۲-۳-۲- گفتار مجزا/ متصل/ پیوسته

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

۲-۳-۳- اندازه‌ی کتاب لغت

تعداد کلمات موجود در کتاب لغت، عامل مهمی در تشخیص کارایی یک سیستم بازشناسی گفتار است. در سیستمهای کتاب لغت کوچک (کمتر از 100 لغت) معمولاً می‌توان به دقتی در حدود 100 % (حتی در سیستم‌های مستقل از گوینده) دست یافت. سیستم‌های با دایره‌ی لغات کوچک در کاربردهایی نظیر تشخیص کارت اعتباری و تشخیص شماره تلفن کاربرد دارند. به هر حال دقت بازشناسی به لغات موجود در کتاب لغت نیز وابسته می‌باشد. اگر کلمات مشابه و گیج کننده باشند، رسیدن به‌دقت 100 % حتی برای کتاب لغت‌های بسیار کوچک، دشوار است.

۲-۳-۴- محدودیت‌های زبانی

محدودیت‌های زبانی[5] را می‌توان با یک مدل از زبان بیان نمود. مدل زبانی یک زبان طبیعی، مرکب از چهار جزء می‌باشد: نمادها، دستور زبان، معنا و جنبه عملی[6]. نمادهای زبان، واحدهای طبیعی هستند که همه‌ی پیغام‌ها از آنها تشکیل می‌گردند و بیانگر کلمات یا واحدهای کوچکتر از کلمه، نظیر هجاها و واج‌ها هستند. دستور زبان، مرکب از محدودیت‌های واژگانی و نحوی است که بیانگر شکل گرفتن کلمات از واحدهای کوچکتر از کلمه و نیز شکل گرفتن جملات از کلمات می‌باشند. جنبه معنایی نیز مرتبط به نحوه‌ی ترکیب کلمات برای شکل‌دادن پیغام‌های با معنی است. به‌عنوان مثال جمله «اسب صحبت می‌کند» از لحاظ نحوی صحیح ولی از لحاظ معنایی نادرست است. در بالاترین سطح، جنبه عملی یک زبان جای دارد که بیانگر وابستگی ادا کردن و معنی کلمه به گوینده‌ها و محیط است. محدودیت‌های

[1] Isolated Word Recognition
[2] Connected Word Recognition
[3] Continues Speech Recognition
[4] Co-Articulation
[5] Linguistic constraints
[6] Pragmatic

معنایی و جنبه عملی به‌ندرت در سیستمهای بازشناسی گفتار استفاده می‌شوند. چرا که این محدودیت‌ها را به‌دشواری می‌توان به‌صورت فرمول بیان کرد. ولی محدودیت‌های دستوری تقریباً در تمامی سیستم‌های بازشناسی گفتار پیوسته به‌صورت محدودیت‌های واژگانی و نحوی مورد استفاده قرار می‌گیرند و تعداد جملات مجاز برای بازشناسی را کاهش می‌دهند.

۲-۳-۵- گفتار مکالمه‌ای

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

۲-۳-۶- محیط

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

۲-۴- اجزای یک سیستم بازشناسی

یک سیستم بازشناسی گفتار شامل اجزای مختلفی می‌شود. در ادامه‌ی این فصل به بررسی اجزای یک سیستم بازشناسی گفتار، از مرحله‌ی نمونه برداری تا مرحله‌ی پردازش زبان می‌پردازیم. بسته به نوع بازشناسی برای گفتار مجزا، متصل یا پیوسته، ممکن است برخی از این اجزاء در سیستم موجود نباشند یا با جزئیات بیشتری در آن سیستم در نظر گرفته شده باشند.

۲-۴-۱- نمونه برداری از سیگنال صوتی

برای استفاده از الگوریتم‌های پردازش سیگنال گسسته، باید موج پیوسته سیگنال گفتار ورودی را به شکل گسسته تبدیل نمود. به‌این منظور باید از موج گفتار ورودی نمونه‌برداری کرد. بنابر قضیه‌ی نایکوئیست، فرکانس نمونه‌برداری موج پیوسته باید حداقل دو برابر بزرگترین مؤلفه‌ی فرکانسی موجود در موج پیوسته باشد. در کاربردهای عملی، موج گفتار پیوسته عموماً از خروجی یک میکروفون و یا از خط

تلفن دریافت می‌شود. بنابراین در بسیاری موارد می‌توان پهنای باند خطوط انتقال تلفن را به‌عنوان معیاری برای تعیین فرکانس نمونه‌برداری در نظر گرفت.

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

۲-۴-۲- استخراج ویژگی از سیگنال گفتار

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

به‌منظور استخراج بردارهای ویژگی باید یک سری پردازش‌ها روی سیگنال انجام شود. این پردازش‌ها عبارتند از: قاب‌بندی[1]، پیش‌تأکید کردن[2]، اعمال پنجره، تبدیل فوریه زمان کوتاه و....

۲-۴-۳- تطبیق الگو

برای ایجاد بانک قواره‌های مرجع در سیستمی که برای بازشناسی گفتار یک گوینده آموزش می‌بینند، از هر الگوی تلفظ شده توسط آن گوینده، یک یا چند قواره بدست آمده و ذخیره می‌شود. الگوی ذخیره شده می‌تواند یک مدل آماری بیانگر مشخصات آن الگو باشد. در هر حال همواره لازم است که میان مجموعه‌ی بردارهای ورودی که هر یک حاوی اطلاعات طیفی اخذ شده از بخشی کوتاه از سیگنال گفتار ورودی هستند و الگوی مرجع، یک تطبیق الگو و مسیریابی زمانی[3] انجام گردد تا بتوان تغییرات سرعت

[1] Framing
[2] Pre-Emphasizing
[3] Time Alignment

بیان عبارت را نیز در سیستم منظور کرد. به این ترتیب امکان محاسبه‌ی یک معیار شباهت مابین کلمه‌ی ورودی و قواره‌های ذخیره شده در سیستم فراهم می‌گردد.

در بعضی از سیستم‌های بازشناسی گفتار بر مبنای کلمه، تطبیق ورودی با قواره‌های مرجع با استفاده از یک روش بهینه‌سازی مشهور به برنامه‌ریزی پویا[1] انجام می‌گردد. حل مسأله‌ی مسیریابی زمانی با روش برنامه‌ریزی پویا به پیچش زمانی پویا با علامت اختصاری DTW معروف است. در این روش برای هر کلمه یا عبارت، یک قواره در حافظه نگهداری می‌شود و هنگام تشخیص الگوی ورودی، میزان انطباق آن با تمامی قواره‌ها بررسی می‌گردد. این کار برای تعداد لغات زیاد، نیازمند حجم حافظه و محاسبات بسیار زیادی است. از طرف دیگر، این روش در بازشناسی گفتار پیوسته کارایی خود را از دست می‌دهد، چرا که مرز دقیق کلمات مشخص نیست. به‌علاوه این روش برای گویندگان متعدد نیز دارای کارایی نیست، زیرا نمی‌تواند تغییرات آکوستیک بین گوینده‌های مختلف را بخوبی مدل کند.

رویکرد دیگر آن است که از هر الگوی لغت‌نامه، یک مدل آماری ساخته شده و هر الگوی لغت‌نامه که مدل آماری آن بیشترین میزان شباهت (احتمال وقوع) را به الگوی ورودی مشاهده داشته باشد، به‌عنوان الگوی ورودی بازشناخته گردد. معمولترین مدل مورد استفاده در این مورد، مدل مخفی مارکوف نام دارد. در مدل مخفی مارکوف، هر واحد از لغت‌نامه توسط مجموعه‌ای از حالتها به‌همراه احتمالات انتقال از حالتی به حالت دیگر نمایش داده می‌شود. در بسیاری از سیستم‌های بازشناسی مبتنی بر مدل مخفی مارکوف، مدل بر پایه‌ی واج‌های تولید شده و سپس در مرحله‌ی نهایی بازشناسی، اطلاعات مربوط به واژگان با اطلاعات واجی ترکیب می‌گردد تا کلمات مورد شناسایی قرار گیرند. مرحله‌ی آموزش مدل مخفی مارکوف نسبت به DTW پیچیده‌تر است و به داده‌های آموزشی بیشتری نیاز دارد. یکی دیگر از شیوه‌های تطبیق الگو استفاده از شبکه‌های عصبی مصنوعی است. شبکه‌های عصبی امکانات خوبی را در جهت پردازش‌های موازی و تطبیق‌یابی در اختیار قرار می‌دهند. در بعضی از تحقیقات انجام شده، روشهای شبکه‌ی عصبی با روشهای دیگری نظیر مدل مخفی مارکوف ترکیب شده‌اند. در این پایان‌نامه از سیستم تشخیص مدل مخفی مارکوف استفاده می‌شود.

۲-۴-۴- پردازش زبان

صرف نظر از آنکه واحد پایه‌ی تشخیص گفتار چیست (کلمه، هجا، آوا یا واج)، برای تعیین چگونگی ادغام این واحدها از نظر ترتیب، متن و معنا، از محدودیت‌های زبان استفاده می‌شود. چنانکه قبلاً گفته شد اجزای مدل یک زبان عبارتند از: نمادها، دستور، معنا و جنبه عملی. این اجزاء به‌عنوان محدودیت‌های

[1] Dynamic Programming

زبان استفاده می‌شوند تا در بازشناسی گفتار، تمامی اطلاعات موجود در ارتباط گفتاری در نظر گرفته شود.

۲-۵- انواع مدل‌سازی در ASR

بیشتر سیستم‌های ASR شامل سه قسمت عمده می‌باشند:

- پردازش سیگنال رو به جلو[1]

این واحد سیگنال گفتار را به دنباله‌ای از بردارهای ویژگی تبدیل می‌کند. استخراج بردارهای ویژگی به‌منظور کلاس‌بندی انجام می‌گیرد. و موجب کاهش نرخ اطلاعات برای پردازش نسبت به سیگنال اصلی می‌شود.

- مدل‌سازی صوتی[2]
- مدل‌سازی زبانی[3]

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

روشهای مختلفی برای بازشناسی گفتار وجود دارد. الگوریتم‌های اصلی بازشناسی گفتار بر پایه روشهای زیر می‌باشند: 1- مدل مخفی مارکوف 2 - پیچش زمانی پویا 3 - شبکه عصبی.

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

در سیستم‌های بازشناسی گفتار که بر اساس مدل مخفی مارکوف می‌باشند، دو قسمت بیشترین زمان بازشناسی را به‌خود اختصاص می‌دهند: یکی«عملیات جستجو» است که بهترین کلمه‌ی منطبق را از روی

[1] Signal processing front-end
[2] Acoustic modeling
[3] Language modeling