بسم الله الرحمن الرحيم

پایان‌نامه کارشناسی ارشد رشته مهندسی صدا

بازشناسی گوینده مبتنی بر روشهای ادغام اطلاعات در سطح تصمیم

دانشجو:

ناهید علینقی زاده

استاد راهنما:

دکتر علی جبار رشیدی

استاد مشاور:

دکتر معصومه شفیعیان

پاییز ۱۳۹۱

چکیده:

در میان تمام فناوریها و روشهای تشخیص هویت زیستی، بازشناسی گوینده بر مبنای اطلاعات صحبت را می‌توان طبیعی‌ترین و اقتصادی‌ترین روش برای سیستمهای ارتباط انسان-ماشین دانست. همچنین توسعه فناوری پردازش صحبت باعث تقویت بسیاری از کاربردهای بازشناسی گوینده شده است. از آنجا که یکی از مهمترین زمینه‌های تحقیقاتی فعال که در سالهای اخیر در بسیاری از کاربردها باعث بهبود عمده در بازشناسی گوینده شده است استفاده از روشها و مدلهای ادغام اطلاعات در سطوح مختلف می‌باشد محور اصلی در این پژوهش، بهبود دقت بازشناسی گوینده با استفاده از ادغام اطلاعات در سطح تصمیم است. در این پایان نامه رویکرد استخراج ویژگیهای جدیدی بر پایه ضرایب کپسترال جهت فراهم سازی منابع تصمیم گیری مورد استفاده در ادغام تصمیم مد نظر قرار گرفته و با توجه به اینکه مشتقات هر تابع بخشی از ویژگیهای مستتر در آن را به نمایش می‌گذارد از مشتق اول و دوم ضرایب کپسترال مل-فرکانس به عنوان بردارهای ویژگی ثانویه استفاده نموده ایم. این رویکرد به مفهوم استفاده و بهره‌گیری همزمان از اطلاعات نهفته در بردار ویژگی، تغییرات(بردار سرعت) و نرخ تغییرات(بردار شتاب) ویژگی می‌باشد. پس از بازشناسی گوینده مبتنی بر این سه بردار ویژگی به صورت مجزا، جهت بهبود دقت و صحت نتایج بازشناسی و شناسایی، اقدام به طراحی چارچوب ادغام اطلاعات در سطح تصمیم نموده‌ایم. استفاده همزمان از این بردارهای ویژگی در بستر ادغام تصمیم تاکنون توسط محققان گزارش نشده است. استفاده از روشهای مناسب جهت خوشه‌بندی بردارهای ویژگی از جمله روش کوانتیزه‌کردن برداری و همچنین توابع تعیین اندازه شباهت از جمله فاصله ماهالانوبیس و فاصله حداکثر شباهت مبتنی بر حد آستانه از دیگر کارهای مهم انجام شده در پژوهش می‌باشد. در مرحله ادغام تصمیم، نتایج بازشناسیهای اولیه را با استفاده از روشهای ادغام تصمیم از جمله روشهای رأی‌گیری، رتبه‌بندی و روش امتیازدهی ترکیب و نتیجه را به عنوان بازشناسی نهایی استفاده ایم. نتایج نشان دهد که روشهای ادغام تصمیم باعث بهبود قابل توجه در دقت بازشناسی گوینده نسبت به حالت بدون ادغام شود. از نتایج دیگر این است که استفاده همزمان از اطلاعات مستتر در بردارهای تغییرات ضرایب کپسترال و بردارهای نرخ این تغییرات برای بازشناسی گوینده باعث بهبود کارایی سیستم بازشناسی گوینده می‌شود.

واژگان کلیدی: بازشناسی گوینده، ادغام تصمیم، استخراج ویژگی، ترکیب ویژگی‌ها، ضرایب کپستروم، فرکانس مل.

فهرست مطالب

عنوان **صفحه**

فهرست جدولها

عنوان **صفحه**

فهرست شکلها

عنوان **صفحه**

۱- فصل اول

کلیات تحقیق

۱-۱-اصول بازشناسی گوینده

بازشناسی گوینده شاخه ای از تشخیص هویت بر پایه ویژگیهای زیستی[1] است که به بازشناسی خودکار شناسه افراد با استفاده از مشخصات ذاتی آنها اطلاق می گردد. تشخیص هویت مبتنی بر معیارهای زیستی برای سیستمهای ارتباط انسان-ماشین در کاربردهایی که در آنها ملاحظات امنیتی وجود دارد جزو مهمترین روشها می باشد. البته غیر از صوت، الگوهای فیزیکی و رفتاری دیگری از قبیل الگوی شبکیه چشم، چهره، اثرانگشت، امضاء و غیره برای تشخیص هویت مبتنی بر سنجش زیستی وجود دارد. در عمل انتخاب یک الگوی زیستی برای بازشناسی الگو حداقل بایستی خواصی همچون مقاوم بودن(در مقابل نویز و سایر شرایط مزاحم)، قابل تشخیص بین افراد مختلف (منحصر به فرد بودن)، در دسترس بودن و قابل قبول بودن را داشته باشد[1].

جدول۱-۱ چهار خاصیت فوق را برای چند الگوی زیستی که مورد استفاده فراوان هستند را مقایسه می کند:

جدول ۱-۱ وضعیت خواص چهارگانه برای الگوهای زیستی

الگوی زیستی	شبکیه	چهره	اثرانگشت	صوت
قابلیت تمیز	بالا	بالا	بالا	قابل قبول
مقاوم بودن	بالا	بالا	قابل قبول	قابل قبول
در دسترس بودن	کم	بالا	قابل قبول	بالا
مقبولیت	قابل قبول	بالا	قابل قبول	بالا

مشخصات فوق را می توان به صورت زیر تعریف کرد:

قابلیت تمیز: یعنی تفاوت مناسب و عمده در الگو برای افراد مختلف باشد.

1-Biometric

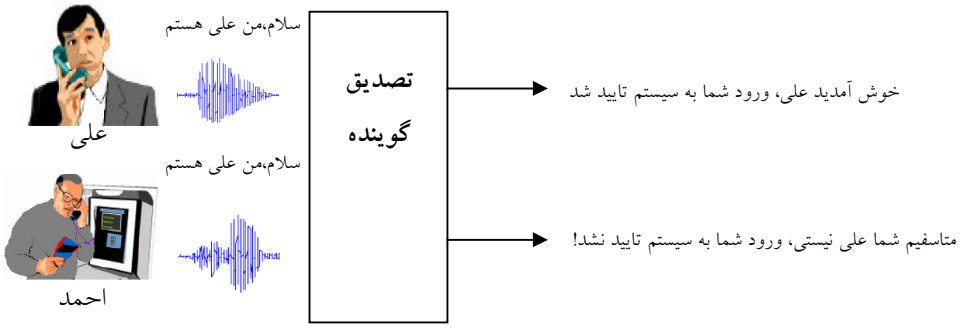
مقاوم بودن: قابل تکرار باشد و در تکرارها تغییرات زیادی در آن رخ ندهد.
در دسترس بودن: به سادگی قابل ارائه به یک حسگر باشد(به سادگی اندازه گیری شود)
قابل قبول بودن: به عنوان یک مشخصه مجاز توسط کاربران قابل مشاهده باشد.
قضاوت در مورد یک الگوی زیستی خوب بسیار پیچیده و وابسته به ویژگیها و شرایط کاربردها می‌باشد.
در میان تمام فناوریها و روشهای تشخیص هویت زیستی، بازشناسی گوینده را می توان طبیعی ترین و اقتصادی ترین روش برای سیستمهای ارتباط انسان-ماشین دانست زیرا اولا وصول اطلاعات صحبت از دیگر الگوها راحتتر است و ثانیا صحبت رویکرد غالب و اصلی تبادل اطلاعات برای انسانهاست و می‌رود که رویکرد غالب برای تبادل اطلاعات انسان-ماشین باشد. همچنین توسعه فناوری پردازش صحبت باعث تقویت بسیاری از کاربردهای بازشناسی گوینده و خصوصا حوزه های زیر شده است:

۱-کنترل دسترسی به تجهیزات فیزیکی یا شبکه های داده
۲-خرید با کارت اعتباری از طریق تلفن یا سایر تراکنشهای بانکی
۳-بازیابی اطلاعات به عنوان مثال اطلاعات مشتری برای مرکز تلفن و فهرست گذاری صوتی
٤-پایش از دور
۵-تطبیق نمونه صوتی در محاکم قضایی

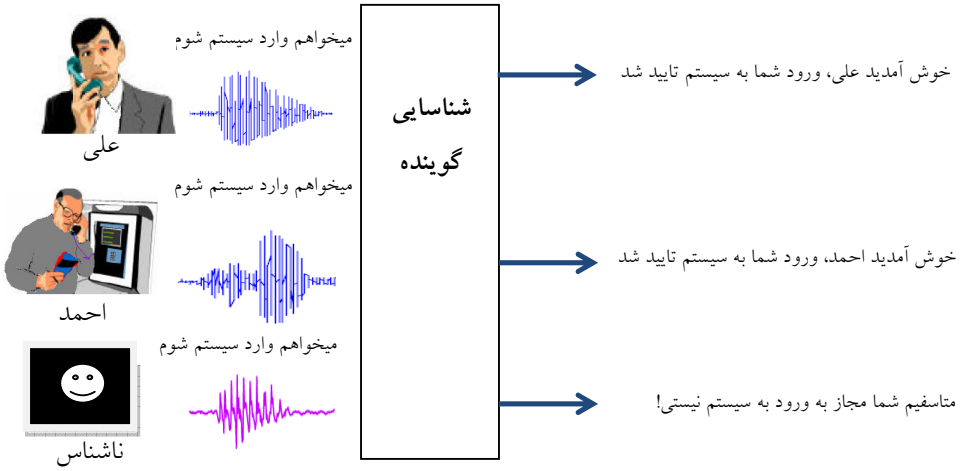
بازشناسی گوینده را می توان به دو وظیفه تصدیق[1] و شناسایی[2] تقسیم کرد. تصدیق عبارت است از تصمیم در مورد اینکه یک صدای نامعین مربوط به یک گوینده ادعایی هست یا نیست. فقط دو تصمیم ممکن است: یا پذیرش اینکه صدا متعلق به گوینده ادعایی است یا رد آن به عنوان یک گوینده متقلب. شناسایی عبارت است از طبقه بندی یک صوت نامعین به عنوان یکی از گویندگان ثبت شده.تعداد تصمیمات ممکن در شناسایی گوینده به تعداد گویندگان ثبت نام شده است و معمولا کارایی سیستم با عکس تعداد گویندگان متناسب است. بنابراین وقتی تعداد گویندگان زیاد است شناسایی بسیار پیچیده تر و مشکلتر از تصدیق گوینده است. این نوع شناسایی که تشریح شد به عنوان شناسایی مجموعه-بسته[3] نیز شناخته میشود. شناسایی مجموعه-باز[4] متناظر با زمانی است که بتوان تصمیم گرفت که صوت نامعین ورودی به هیچکدام از گوینده های ثبت شده متعلق نباشد.

1- Verification
2- Identification
3- Closed-Set
4- Open-Set

بنابراین تعداد تصمیمات ممکن از تعداد گویندگان ثبت شده یکی بیشتر است. شکلهای ۱-۱ و ۱-۲ به طور کلی نحوه کار سیستمهای تصدیق و شناسایی گوینده را نشان می دهند.



شکل ۱-۱ سیستم تصدیق گوینده



شکل ۱-۲ سیستم شناسایی گوینده

شناسایی مجموعه-باز ترکیب شناسایی مجموعه بسته و تصدیق گوینده است. همچنین بازشناسی گوینده را میتوان به بازشناسی وابسته به متن و بازشناسی مستقل از متن تقسیم نمود. در بازشناسی وابسته به متن، سیستم دقیقا متن صحبت را می داند. در بازشناسی مستقل از متن سیستم از متن صحبت آگاهی ندارد. با داشتن اطلاعات متن صحبت، سیستم می تواند گوینده خاص را با اجزاء کلمه یا حروف معین شناسایی کند و بنابراین سیستمهای وابسته به متن معمولا کارایی بهتری نسبت به سیستمهای مستقل از متن دارند اما در این سیستمها نیاز به همکاری و تعامل زیاد گوینده می باشد

و بنابراین برای کاربردهایی که کنترل قوی روی ورودی کاربر وجود دارد بکار می رود.
کاربرد سیستم مستقل از متن بیشتر از سیستم وابسته به متن بوده و کاربرپسندتر است اما بدون اطلاعاتی از متن صحبت رسیدن به کارایی بالا دشوار است.در کاربردهای مستقل از متن، استفاده از یک تشخیصگر متن که بتواند از متن صحبت اطلاعاتی را فراهم کند می تواند دقت بازشناسی گوینده را بهبود دهد[2]. هر چند سیستمهای مستقل از متن به عنوان بستر بهتری برای ارائه و ارزیابی فناوریها و سیستمهای بازشناسی گوینده پذیرفته شده اند، اما بسیاری از کاربردهای تجاری و صنعتی توجه بیشتری به بازشناسی وابسته به متن یا بازشناسی با متن محدود[1] نموده اند.

۲-۱-مروری مختصر بر تاریخچه بازشناسی گوینده

پژوهش در زمینه بازشناسی گوینده حدود ۵۰ سال پیشینه دارد و همواره به عنوان حوزه ای فعال در پردازش زبان مطرح می باشد. توسعه روشها و فناوری بازشناسی گوینده همواره رابطه نزدیکی با پیشرفت در پردازش سیگنال و صحبت و فناوری رایانه داشته و دارد. بازشناسی گوینده توسط انسان به طور وسیع از دهه ۱۹۶۰ میلادی مطالعه و بررسی شد. انگیزه و هدف این مطالعات، فهم چگونگی بازشناسی گوینده ها توسط انسان و قابلیت اطمینان انسان در بازشناسی یک گوینده بود[3]. یکی از مهمترین کارهای پژوهشی که باعث گسترش تحقیقات در بازشناسی گوینده با استفاده از رایانه گردید مطرح کردن نمودار طیفی[2] توسط کرستا[3] به عنوان یک ابزار شناسایی شخص بود که در سال ۱۹۶۲ در نشریه نیچر منتشر گردید.
در دهه ۱۹۷۰ توجهات به بازشناسی گوینده با استفاده از رایانه و آغاز بازشناسی خودکار گوینده[4] (ASR) بود. سیستمهای بازشناسی گوینده در این دوره از پژوهشها محدود به تعداد محدودی گوینده(کمتر از ۲۰ نفر) بود[4]. روشها و فنون تبدیل فوریه، پیشگویی خطی[5]، آنالیز کپسترال[6] در این دوره برای استخراج پارامترها و ویژگیها بکار برده شدند. متوسط زمان طولانی این پارامترها به عنوان مرجع گوینده استفاده می شد.
در دهه ۱۹۸۰ روشهای آماری پیچیده تر بازشناسی الگو از قبیل همپیچی زمانی پویا[7] [5] و کمی سازی برداری[8] [6] برای سیستمهای در مقیاس بزرگ بازشناسی گوینده(بیشتر از ۱۰۰ نفر) ابداع و

1 - Text-Constraint
2 - Spectrogram
3 - L. G. Kersta
4 - Automatic Speaker Recognition(ASR)
5 - Linear Predictive
6 - Cepstral Analysis
7 - Dynamic Time Warping (DTW)
8 - Vector Quantization (VQ)

استفاده شد. همچنین استفاده توأم از ویژگیهای دینامیک و استاتیک برای بازشناسی گوینده نیز در همین دهه توسط محققان انجام پذیرفت[7].

از سالهای حدود ۱۹۹۰ در دسترس بودن پایگاه های بزرگ داده صحبت مانند YOHO مطالعات روی مدلهای پیچیده تر برای بازشناسی گوینده را تقویت کرد. این مدلها شامل مدلهای تصادفی[۱] (از قبیل مدل مخفی مارکف[۲] [8] و مدل ترکیبی گوس[۳])، شبکه های عصبی (مانند پرسپترون چندلایه ای[۴] [9] و توابع پایه شعاعی[۵] [10]) و ماشینهای بردار پشتیبان[۶] [11] و دیگر مدلها می باشد. در بین این روشهای مدلسازی، روش GMM به عنوان مؤثرترین روش در تعیین توزیع چگالی داده های صحبت شناخته می شود و روش غالب در مدلسازی برای بازشناسی گوینده است.

برای استخراج ویژگیها، ضرایب کپسترال که برگرفته از مدل شنوایی انسان است یعنی ضرایب کپسترال مل- فرکانس[۷] و ضرایب دینامیک آنها به عنوان روش غالب در استخراج پارامترهای ویژگی شناخته شده می باشند. از طرفی محققان روشهای مختلف نرمال سازی امتیاز[۸] برای بازشناسی مقاوم گوینده را ارائه نموده اند[12].

۳-۱-مشکلات و چالشهای روشها و فناوریهای موجود بازشناسی صحبت

سیستمها و روشهای موجود بازشناسی صحبت در آزمایشگاه یا بعضی کاربردهای خاص با شرایط کاری و آموزش سطح بالا و طراحی ماهرانه، خیلی خوب کار می کنند.

نتایج تجربی نشان می دهد که بازشناسی خودکار گوینده در چنین محیطهای ایده آلی به خوبی بازشناسی توسط انسان است. اما از منظر عملی و کاربردی، کارایی سیستمهای بازشناسی گوینده فعلی در کاربردهای واقعی هنوز از درجه مقاوم بودن و قابلیت اطمینان مناسبی در مقایسه با کارایی بازشناسی توسط انسان برخوردار نیستند. اولین مسئله پیش روی سیستمهای بازشناسی گوینده، بهبود سطح مقاوم بودن سیستم تحت شرایط واقعی و غیرمنطبق است. برای بازشناسی گوینده، عدم تطبیقها اصولا به دو دلیل رخ می دهد:

۱-تغییرات فردی گوینده مربوط به شیوه صحبت کردن.

---

1 - Stochastic Models
2 - Hidden Markov Model(HMM)
3 - Gaussian Mixture Model (GMM)
4 - Multilayer Perceptron (MLP)
5 - Radial Basis Function (RBF)
6 - Support Vector Machines (SVM)
7 - Mel-Frequency Cepstral Coefficients (MFCC)
8 - Score Normalization

۲-تغییرات آکوستیکی محیط

سیستم صوتی انسان در گفتن یک عبارت ثابت در زمانهای مختلف نیز دارای تغییرات است و این باعث می شود که تطبیق گفته های یک فرد با ویژگیهای خودش به خوبی انجام نشود و یکی از منابع اصلی خطا در سیستمهای بازشناسی گوینده همین تغییرات ذاتی مربوط به گوینده است. تغییرات آکوستیکی محیط نیز بدلیل اعوجاجات و اختلالات غیرقابل پیش بینی مختلف در حین وصول و انتقال داده ها می باشد. به عنوان مثال در کاربردهای بازشناسی گوینده تلفنی مانند سیستمهای تراکنش تلفنی بانک، داده های صوتی ممکن است در محیطهای مختلف، با تلفنهای متفاوت و از طریق کانالهای ارتباطی مختلف دریافت شود. نویز زمینه و اعوجاج کانال و گوشی باعث تغییر در ساختار طیفی داده های صوتی می شود و ویژگیهای مستخرجه آکوستیکی (مثلا پارامترهای MFCC) نمی تواند اطلاعات گوینده را به درستی در بر داشته باشد.

برای بهبود کارایی سیستمهای بازشناسی گوینده در محیطهای واقعی نیز محققان فعالیتهای متنوعی را انجام داده اند. از جمله این کارها می توان به تحقیقات در زمینه استخراج ویژگیهای سطح بالا و جدید از قبیل ویژگیهای عروضی، لغوی و شیوه صحبت کردن و امثال آن به عنوان مکمل ویژگیهای آکوستیکی سطح پایین(مانند ضرایب کپسترال) برای بازشناسی مقاوم گوینده اشاره نمود که یکی از زمینه های تحقیقاتی بسیار فعال در این حوزه می باشد[13].

به هر حال استخراج ویژگی مؤثر و مدل سازی مناسب داده ها هنوز به خوبی کاربردی نشده و فقط بخشی از مشکلات بازشناسی را حل می کند و همیشه نمی توان از آنها استفاده نمود. یکی از زمینه هایی که در سالهای اخیر در بسیاری از کاربردها باعث بهبود عمده نتایج شده است استفاده از روشها و مدلهای ادغام اطلاعات[1] است که در سطوح مختلف قابل کاربرد و به عنوان یکی از مهمترین زمینه های تحقیقاتی فعال مطرح است[14].

البته یکی دیگر از رویکردهای بازشناسی گوینده استفاده از الگوهای چندگانه به صورت همزمان و بهره گیری از ادغام اطلاعات چندحسگری[2] است مثلا ادغام الگوهای چهره، اثرانگشت و صوت از جمله این رویکردهاست. پیچیدگی دریافت اطلاعات کاربر و هزینه ها در سیستم چند حسگری از جمله موانع توسعه این رویکردها می باشد.

1 - Data Fusion
2 - Multisensor Data Fusion

۳-۱-مسئله تحقیق و هدف آن

بهبود دقت بازشناسی گوینده با استفاده از ادغام اطلاعات در سطح تصمیم مورد توجه این تحقیق است. همچنین استخراج ویژگیهای جدیدی بر پایه ضرایب کپسترال جهت فراهم سازی منابع تصمیم گیری مورد استفاده در ادغام تصمیم از دیگر رویکردهای این پایان نامه می باشد. با توجه به اینکه مشتقات هر تابع بخشی از ویژگیهای مستتر در آن را به نمایش می گذارد که انتظار داریم مشتق اول و دوم ضرایب کپسترال مل-فرکانس و مشتق اول و دوم خود کپسترال مل-فرکانس به عنوان بردارهای ویژگی جدید در ترکیب با هم و با استفاده از روشهای ادغام تصمیم باعث بهبود مناسبی در دقت بازشناسی گوینده نسبت به حالت بدون ادغام شود. این رویکرد به مفهوم استفاده و بهره گیری همزمان از اطلاعات نهفته در بردار ویژگی، تغییرات(بردار سرعت) و نرخ تغییرات(بردار شتاب) ویژگی می باشد. رویکرد استفاده همزمان از این بردارهای ویژگی در بستر ادغام تصمیم تاکنون توسط محققان گزارش نشده است.

هدف این تحقیق بهبود کارایی بازشناسی گوینده با استفاده از ادغام همزمان اطلاعات حاصل از بازشناسی گوینده توسط هرکدام از بردارهای ویژگی فوق در سطح تصمیم است. در این تحقیق سعی می کنیم که به سؤالات زیر پاسخ دهیم:

آیا استفاده از اطلاعات مستتر در بردارهای تغییرات کپسترال و ضرایب کپسترال و بردارهای نرخ این تغییرات برای بازشناسی گوینده مفید است؟

چگونه به طور مؤثر می توان اطلاعات گوینده را از تغییرات و نرخ تغییرات بردارهای ویژگی کپسترال و ضرایب کپسترال استخراج کرد؟

چگونه می توان از مزایای ادغام اطلاعات در سطح تصمیم بر مبنای ادغام ویژگیها و تغییرات و نرخ تغییرات آنها برای بازشناسی گوینده بهره برد؟

٤-۱-مفاهیم تحقیق

در این پایان نامه بطور مقدماتی در خصوص مفهوم و کاربردهای بازشناسی گوینده و تاریخچه مختصر آن پرداختیم. یکی از حوزه‌های تحقیقاتی مورد توجه محققان، ارائه الگوهای جدید به منظور افزایش دقت عملکرد سیستم‌های بازشناسی می‌باشد. از جمله روشها و فناوریهای مؤثر در بهبود