سومین کنفرانس ملی مباحث نوین در کامپیوتر و فناوری اطلاعات
3nd National Conference on Advanced Topics in Computer and Information Technology
بیست و هشتم آذر ماه ۱۳۹۸

بهینه سازی سیستم همیار نابینایان و کم بینایان مبتنی بر گفتار

حامد مراونی
گروه مهندسی کامپیوتر - دانشگاه آزاد اسلامی واحد ماهشهر
Hamed.maravani57@gmail.com
مرجان عبد یزدان
گروه مهندسی کامپیوتر - دانشگاه آزاد اسلامی واحد ماهشهر
Abdeyazdan87@yahoo.com

چکیده

سیستم همیار نابینایان از دو بخش شناسایی گفتار و ناوبری بر روی نقشه تشکیل شده است. دربخش گفتار سیگنالهای ورودی از طریق الگوریتم ADT بخش‌بندی شده و سپس ویژگی MFCC از آنها استخراج می‌شود. ویژگی‌های حاصل، از طریق روش تطبیق DTW با هم مقایسه می‌شوند. سیستم علاوه بر ۴ جهت اصلی، با ذخیره‌سازی ویژگی‌های عناوین اماکن مختلف (حداکثر ۵ نام متفاوت) نیز آموزش داده شده است. کارایی عملیات جداسازی مطلوب و در حدود ۹۰ درصد است. در بخش ناوبری بر روی نقشه، از پایگاه داده، موقعیت جغرافیایی مکان مورد نظر بازیابی شده و با توجه به نقطه فعلی شخص نابینا، مسیر بخش‌بندی می‌شود. سپس، سیستم، هر بخش را از طریق فایل صوتی مرتبط قرائت می‌کند. همچین در سیستم یک شبیه‌ساز GPS پیاده‌سازی شده که از طریق آن می‌توان سیستم را به صورت مجازی تست کرد.

کلید واژگان: پردازش گفتار ،سیستم همیار نابینایان، MFCC، DTW

۱-مقدمه

هرآنچه مربوط به تعامل انسان و کامپیوتر است HCI تلقی می‌شود، از سخت‌افزار گرفته تا نرم‌افزارهای مختلف که به نکات طراحی، ارزیابی و پیاده‌سازی سیستم‌های کامپیوتری تعاملی برای استفاده انسان می‌پردازند. می‌توان گفت از زمان ظهور کامپیوتر مفهوم تعامل کامپیوتر و انسان نیز مطرح شده است ؛ زیرا یک کامپیوتر پیچیده اگر نتواند توسط انسان مورد استفاده قرار گیرد، سودی نخواهد داشت. تعامل بین کامپیوتر و انسان در طی سال‌ها نه تنها از لحاظ کیفی تغییر نموده بلکه شاخه‌های متعددی نیز در آن به وجود آمده است، به عنوان مثال می‌توان به استفاده از واسطهای کاربری هوشمند به جای واسطهای command/action اشاره نمود، در راستای این هوشمندی مفاهیم عامل نیز مطرح می‌گردد.

خصوصیت خودمختار بودن عامل‌ها، تعامل انسان و کامپیوتر را از یک حالت صریح به یک حالت ضمنی تبدیل نموده است. اگر نیاز باشد کاربر به صورت مستقیم و صریح با سیستم ارتباط برقرار کند باید بسیاری از کارها را خودش آغاز نماید و به بسیاری از رویدادها نظارت داشته باشد. اگر کاربران غیرحرفه‌ای بوده و آموزش‌های لازم را ندیده باشند، باید این روش مستقیم را تغییر داد تا آنها نیز بتوانند به صورت کارا از سیستم استفاده نمایند. با به وجود آمدن عامل‌ها، این نوع تعامل کاربران با سیستم تغییر یافته و یک روش غیرمستقیم ایجاد شده است. نمی‌توان گفت عامل‌ها راهی غیر از ارتباط مستقیم برای کاربر هستند بلکه کار آنها بهبود بخشیدن به توانایی‌های کاربر برای ارتباط مستقیم با واسطهای کاربری است، به عبارت دیگر عامل‌ها به عنوان کمک‌دهنده در واسطهای کاربری استفاده می‌شوند. ما عامل‌ها را به عنوان افراد هوشمند و خبره‌ای می‌بینیم که در پس واسطهای کاربری کمک می‌نمایند تا کاربران کارهای خود را به نحو بهتری انجام دهند.

آنچه در طراحی HCI مهم است فراهم کردن سهولت استفاده، مطلوب بودن و رضایت خاطر برای کاربران می‌باشد. برای تحقق این اهداف واسطهای کاربری روزبه‌روز به دنیای واقعی نزدیک‌تر گشتند. واسطهای کاربری نقش عمده‌ای در نمایش اطلاعات به کاربران ایفا می‌نمایند و با وجود رشد روز افزون حجم اطلاعات در شبکه جهانی اینترنت، نمایش این اطلاعات به کاربر به گونه‌ای که رضایت خاطر وی را جلب نماید، مسئله بسیار با اهمیتی است. حال اگر این کاربران دارای معلولیت‌هایی مانند نابینایی باشند، طراحی واسطهای کاربری می‌بایست با در نظر گرفتن این ناتوانی صورت گیرد. متاسفانه اغلب نرم افزارها و قطعات کامپیوتری با این فرض طراحی شده‌اند که تمامی کاربران از حواس پنجگانه برخوردارند و کمتر به طراحی واسطهای کاربری برای کاربران خاصی که تعدادشان کم هم نیست، پرداخته شده است. هدف ما بهینه سازی یک واسط کاربری هوشمند است که کاربر نیازی به دیدن آن نداشته باشد و با استفاده از فرامین صوتی بتواند از خدمات آن استفاده نماید.[۱].

۱-۱- بیان مسئله

اولین گامی که در این پروژه برداشته شد، آشنایی بیشتر با کاربران مورد نظر و بررسی نیازمندی‌ها، امکانات و مشکلات آنها بوده تا بتوانیم سیستمی مناسب با نیازهای این گروه از افراد ارائه نماییم. برای این منظور با مراجعه به روزنامه ایران سپید (روزنامه‌ای با خط بریل مخصوص افراد نابینا) و مصاحبه با روزنامه‌نگاران نابینای این روزنامه توانستیم با اساتید نابینای کامپیوتر در موسسه عصای سفید آشنا شویم

و از نزدیک چگونگی استفاده این افراد از کامپیوتر و انجام کارهای روزمره‌شان را شاهد باشیم و با نرم‌افزارهایی که این افراد به عنوان ابزارهای کمکی در کار با کامپیوتر استفاده می‌نمایند، آشنا گردیم. طی این مصاحبه‌های حضوری ایده‌های پروژه و شرح مسئله را با آنها در میان گذاشتیم و بازخوردهای قابل ملاحظه‌ای دریافت کردیم که در ایجاد بستر مناسب برای ادامه ارتباط به منظور طراحی مناسب‌تر و همچنین مرحله تست سیستم پیاده‌سازی شده، بسیار ارزشمند است.

در ابتدای کار مطالعات خود را بر مبنای بررسی سیستم-های طراحی شده برای نابینایان قرار دادیم که در آنها بهبود تعامل فرد نابینا با کامپیوتر مد نظر قرار گرفته بود، برطبق این مطالعات توانستیم دسته‌بندی شکل ۱-۱ را برای سیستم‌هایی که تاکنون با هدف نابینایان طرح‌ریزی شده‌اند، ارائه دهیم.

۱-۲-اهمیت پژوهش

ممکن است طراحی سخت‌افزارهای جانبی برای تسهیل کار این کاربران با نرم افزارها قدم مهمی در فراهم نمودن امکان استفاده از خدمات برای نابینایان برداشته باشد اما نباید این مسئله را نادیده گرفت که عمدتا این وسایل جانبی با صرف هزینه‌های بالایی همراه بوده و تهیه و استفاده از آنها برای تمامی کاربران امکان‌پذیر نمی‌باشد. لااقل مطالعات و گفتگوهایی که در ایران با این افراد انجام دادیم ما را به این نتیجه رسانید که این سخت‌افزارها به دلایلی از جمله بالا بودن قیمت‌ها، به ندرت و آن هم در مکان‌های خاص و برای کاربردهای خاصی مانند نابینایانی که در دفاتر روزنامه‌نگاری مشغول به کار هستند، استفاده می‌شود و بقیه از آن محروم می‌باشند.

براساس داده‌های به دست آمده از [۲] و [۳]، ۳۱۴ میلیون نفر از جمعیت دنیا را افرادی با مشکلات بینایی تشکیل می‌دهند، یعنی ۴ درصد مردم جهان، که ۴۵ میلیون نفر از آنها (۱۴ درصد) نابینا هستند، این در حالیست که ۸۷ درصد آنها در کشورهای در حال توسعه زندگی می‌کنند و از این میان جمعیت ایران سهم بالایی دارد.

با وجود مزایای متعددی که تکنولوژی‌های دیجیتالی در مقابل روش‌های قدیمی مانند کتاب‌های کاغذی در اختیار کاربران نابینا قرار می‌دهند اما همچنان این کاربران در استفاده از خدمات کامپیوتری با مشکلاتی مواجه هستند. حجم وسیعی از کنترل‌های بصری موجود در صفحات وب و تلفن های هوشمند، کار صفحه‌خوان‌ها را برای تفسیر این صفحات مشکل نموده‌اند [۴]، صفحه‌خوان‌های مورد استفاده این کاربران به صورت سلسله‌مراتبی و از ابتدای یک صفحه خط به خط شروع به خواندن آن نموده و اطلاعات ظاهری صفحه و چیدمان مولفه‌ها و فرمت آنها را منتقل نمی‌کنند.

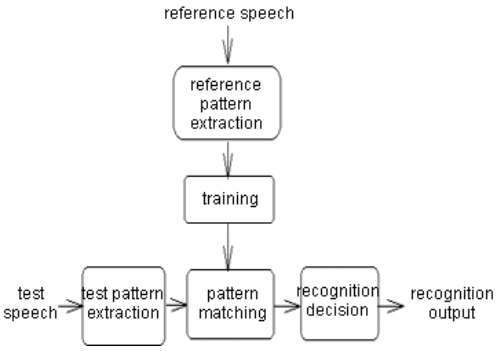
۲-کلیات تحقیق

۲-۱-گفتار چیست

هنگامی که ما انسانها صحبت می‌کنیم، اجازه می‌دهیم هوا از ریه ما، از طریق دهان و حفره بینی عبور کند و این جریان هوای محدود شده، با زبان و لبهای ما تغییر می‌کند. این انقباض و انبساط هوا، یک موج صوتی تولید می‌کند [۱۳].

۲-۲-بازشناسی گفتار

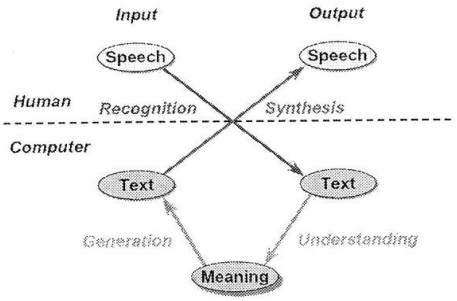
فناوری بازشناسی گفتار(تشخیص گفتار) نرم‌افزاری است که قادر است صوت را به متن تبدیل کند. فناوری تشخیص گفتار به رایانه‌ای که توانایی دریافت صدا را دارد برای مثال به یک میکروفون مجهز است، این قابلیت را می‌دهد که صحبت کاربر را متوجه شود. این فناوری در تبدیل گفتار به متن و یا به عنوان جایگزینی برای ارتباط با رایانه کاربرد دارد. برقراری ارتباط گفتاری با رایانه‌ها به‌جای استفاده از صفحه‌کلید و ماوس یکی از زمینه‌های تحقیقاتی مهم چند دهه اخیر است و شرکت‌های بزرگی چون مایکروسافت، فیلیپس، ای ال ای‌اتی، ای‌بی‌ام، سالانه هزینه‌های هنگفتی را برای این منظور پرداخت کرده و می‌کنند.



شکل ۱-۱ : دیاگرام یک سیستم تشخیص گفتار

تشخیص (بازشناسی) گفتار یا به طور عمومی بازشناسی خودکار گفتار (ASR)، عمل شناسایی گفتار انسان توسط

کامپیوتر است. تعریف فنی ژورافسکی [۱۲] بیان می‌دارد: سیستم ASR، سیستمی است برای نگاشت سیگنالهای صوتی به رشته‌ای از کلمات. در واقع سیستم ASR، سیگنال گفتار را به کلمات تبدیل می‌کند. شکل۱-۲ الگوریتم پردازش گفتار انسان و کامپیوتر را نشان می‌دهد.



شکل ۱-۲ : پردازش گفتار در انسان و کامپیوتر

برای مثال سیستم تلفن گویای بانک را در نظر بگیرید. با بانک خود تماس می‌گیرید. صدای یک نوار را می‌شنوید: « برای اطلاع از میزان پول در حساب کلید ۱، برای تغییر رمز کلید ۲ و ...» شما هم از این‌که می‌توانید از فناوری روز استفاده کنید خوشحال می‌شوید و اطلاعات موردنظر را کسب می‌کنید؛ اما آیا می‌دانید که در بعضی کشورها برای ارتباطات این‌چنینی از فشردن دکمه‌های تلفن استفاده نمی‌کنند؟ در این کشورها افراد حتی زحمت چنین کاری را به خود نمی‌دهند بلکه به‌راحتی منظور خود را می‌گویند و سیستم کار موردنظر آنها را انجام می‌دهد. درواقع با استفاده از سیستم تشخیص گفتار، این اتفاق روی می‌دهد. لابد می‌گویید این موضوع تازه‌ای نیست و مدت‌هاست که این حرف‌ها زده می‌شود. حق با شماست. بیش از یک دهه است که محققان سعی می‌کنند تا صوت را به‌عنوان یک ورودی برای رایانه تعریف کنند. حالا به نظر می‌رسد این تحقیقات به نتایج خوبی رسیده‌اند. طی این سال‌ها تلاش زیادی روی تشخیص گفتار صورت گرفته است. اما با توجه به عوامل زیادی که در این الگوریتمها مؤثر هستند، همواره عملیات تشخیص با خطا روبه‌رو بوده است. تارهای صوتی انسان خصوصیاتی غیرخطی دارند و از طرف دیگر عوامل مختلفی از جنسیت تا حالت عاطفی فرد در فعالیت آنها تأثیرگذار است. درنتیجه تلفظ صوتی می‌تواند به لهجه، طرز تلفظ، طرز گفتار و میزان شمرده بودن آن، درشتی صدا، تودماغی حرف زدن، زیروبمی صدا، درجه صدا (بلندی) و سرعت ادای کلمات بستگی داشته باشد. علاوه بر این‌ها از آنجا که معمولاً افراد در محیطی صحبت می‌کنند که صداهای محیطی نیز وجود دارد، این مسئله پیچیده‌تر می‌شود؛ به شکلی که تشخیص گفتار حتی از تولید گفتار سخت‌تر و پیچیده‌تر است. دقت یک سیستم تشخیص گفتار بستگی به شرایط آزمودن دارد. در شرایط محیطی و گفتاری خاص یک سیستم بسیار خوب عمل می‌کند اما در شرایط عمومی این دقت کاهش می‌یابد. این شرایط ابعاد گوناگونی دارند که می‌توان به‌اختصار به بعضی از آن‌ها اشاره کرد.

۲-۳-کاربردهای تشخیص گفتار

فناوری بازشناسی گفتار، بر پایه این ویژگی‌ها در طیف گسترده‌ای از محصولات قابل‌استفاده است. نمونه‌هایی از زمینه‌های کاربرد آن عبارت‌اند از: خودروها، لوازم‌خانگی الکتریکی و الکترونیکی، اسباب‌بازی‌ها، عروسک‌ها و سرگرمی‌های رایانه‌ای، سیستم‌های دستیار افراد کم‌توان و سالخورده، نرم‌افزارهای رایانه‌ای مدیریتی، سیستم‌های آموزش زبان یا به‌عنوان نمونه دادن فرمان‌های صوتی به خودرو هنگامی‌که راننده مشغول رانندگی است و نمی‌تواند کار دیگری انجام دهد، این فرمان‌ها می‌تواند شامل موارد ذیل باشد: تنظیم آینه‌های بغل و عقب، کنترل بالابر شیشه‌ها، کنترل قفل کودک، کنترل روغن ترمز و موتور یا بنزین در حال حرکت، کنترل رادیو یا هر نوع رسانه دیگر در خودرو، کنترل برف‌پاک‌کن‌ها، تنظیم صندلی‌ها، کنترل چراغ‌ها و هر نوع دستور دیگری که انجام آن نیازمند حرکت اضافی راننده و یا سرنشینان است. چنین نرم‌افزاری موجود است و به‌خوبی در محیط پر نویز، عمل می‌کند مثلاً در خودرویی با سرعت ۱۰۰ کیلومتر در ساعت با شیشه‌های باز و در بزرگراه آزموده شده و پاسخ مناسب گرفته است.

در ادامه از سیستم‌هایی که از تشخیص گفتار استفاده کرده‌اند، برای نمونه مثال‌هایی می‌آوریم: به‌عنوان یک کاربر رایانه، احتمالاً باقابلیت گفتاری مجموعه آفیس به‌عنوان یکی از ویژگی‌های جذاب و تا حدی فانتزی آن برخورد کرده و یا با آن کار کرده‌اید. به کمک این قابلیت شما به‌جای استفاده از صفحه‌کلید برای تایپ مطالبتان به‌راحتی می‌توانید با خواندن متن موردنظرتان و انتقال گفتارتان به کمک یک میکروفن معمولی به رایانه مطلب موردنظرتان را

تایپ‌شده تحویل بگیرید. حتی برای ذخیره کردن، کپی کردن، گذاشتن عکس در متن و ..... به‌جای کلیک‌های پشت سر هم و گاهی با تعداد بالا، می‌توانید فرمان مربوطه را به کمک گفتار به نرم‌افزار بدهید تا کار شما را انجام دهد. جدای از این که توانایی درست‌کار کردن این قابلیت آفیس چه قدر است، یک محدودیت بزرگ در سر راه استفاده از آن برای ما ایرانیان وجود دارد، این قابلیت فقط با زبان انگلیسی سازگار است.

کاربردهای نیازمند پردازش صحبت اغلب در دو دسته ترکیب صحبت و تشخیص صحبت مورد بررسی قرار می‌گیرند. ترکیب صحبت عبارت است از فن آوری تولید مصنوعی صحبت به وسیله ماشین و به طور عمده از پرونده‌های متنی به عنوان ورودی آن استفاده می‌گردد. در اینجا باید به یک نکته مهم اشاره شود که بسیاری از تولیدات تجاری که صدای شبیه به صحبت انسان ایجاد می‌کنند، در واقع ترکیب صحبت انجام نمی‌دهند بلکه تنها یک تکه صدای انسان را که به صورت دیجیتال ضبط شده، پخش می‌کنند. این روش کیفیت صدای بالایی ایجاد می‌کند اما به واژه‌ها و عبارات از پیش ضبط شده محدود است. از کاربردهای عمده ترکیب صحبت می‌توان به ایجاد ابزارهایی برای افراد دارای ناتوانی بینایی برای مطلع شدن از آنچه بر روی صفحه کامپیوتر می‌گذرد، اشاره کرد. تشخیص صحبت عبارت است از تشخیص کامپیوتری صحبت تولید شده توسط انسان و تبدیل آن به یک سری فرامین یا پرونده‌های متنی. کاربردهای عمده موجود برای این گونه سیستمها دربرگیرنده بازه گسترده‌ای از سیستمهاست. از سیستمهای دیکته کامپیوتری گرفته تا سیستمهای کنترل کامپیوترها به وسیله صحبت و به طور خاص سیستمهای فراهم آورنده امکان کنترل کامپیوترها برای افراد ناتوان از لحاظ بینایی یا حرکتی. کاربرد مورد نظر ما از لحاظ نحوه پیاده سازی و استفاده، تناسب فراوانی با خانواده دوم یعنی تشخیص کامپیوتری صحبت دارد، ولی از لحاظ اهداف و کاربردها می‌تواند در خانواده‌ای جداگانه از کاربردهای نیازمند پردازش صحبت قرار گیرد. ترکیب و تشخیص کامپیوتری صحبت مسائل دشواری هستند. روشهای مختلفی مورد آزمایش قرار گرفته‌اند که موفقیت کمی داشته‌اند. این زمینه از زمینه‌های فعال در تحقیقات پردازش سیگنال دیجیتال (دی. اس. پی) بوده و بدون شک سالها این گونه خواهد ماند. در حال حاضر از ابزارهای برنامه نویسی جا افتاده در زمینه‌های برشمرده شده، می‌توان به ای. پی. آی صحبت شرکت مایکروسافت اشاره نمود که دارای تواناییهای عمده‌ای در زمینه‌های تشخیص و ترکیب صحبت است و توانایی آن تا حدی گسترده است که در یک محصول بزرگ و عملی از آن استفاده شده است.

۲-۴-مشکلات و عوامل موثر در بازشناسی گفتار

در یک سیستم بازشناسی گفتار، پارامترهای مختلفی تعیین کننده درجه توانایی سیستم هستند:

۲-۴-۱- مقایسه دریافت انسان از گفتار با ASR

انسانها هنگامی که گوش می‌کنند، علاوه بر گوش، از شناختی که درباره گوینده ه و موضوع دارند نیز استفاده می‌کنند. کلمات به طور دلخواه دنبال هم قرار نگرفته‌اند، آنها یک ساختار دستوری دارند که اشخاص از آن برای پیشگویی کلمات گفته نشده استفاده می‌کنند. به علاوه اصطلاحات و تکه‌کلام‌ها نیز پیشگویی را ساده‌تر می سازد. اما در ASR ما تنها سیگنال گفتار را داریم. البته ما مدلی برای ساختار دستوری ایجاد می‌کنیم و از برخی انواع مدلهای آماری برای بهبود پیشگویی استفاده می‌کنیم اما هنوز مشکل این که چگونه دانش جهانی و دانش گوینده را مدل کنیم، وجود دارد. ما می‌توانیم دانش جهان را به طور جامع مدل کنیم، اما سوال این است که واقعا چه قدر لازم است که ASR شبیه به دریافت انسانها باشد؟

۲-۴-۲- زبان بدن

افراد گوینده، تنها با گفتار ارتباط برقرار نمی‌کنند. آنها از حرکات بدن، حرکات دست، تغییرات چشم و حالتها نیز استفاده می‌کنند. این اطلاعات کاملا در ASR گم شده است. این مساله در محدوده تحقیقات چند کیفیتی است که بررسی می‌کند که چگونه زبان بدن را برای پیشرفت ارتباط انسان و کامپیوتر مورد توجه قرار دهیم.

۲-۴-۳- نویز

گفتار اغلب در محیطهای پر سر و صدا تولید می‌شود: ساعت تیک تاک می‌کند، کامپیوتر صدا می‌دهد، رادیو روشن است، گویندگان دیگری نیز وجود دارند. به این اطلاعات ناخواسته (صداهای ناخواسته) در سیگنال گفتار

نویز گفته می‌شود.

در ASR می‌خواهیم این نویزها را تشخیص دهیم و از سیگنال گفتار حذف کنیم. نوع دیگری از نویزها، اثر پژواک است. پژواک سیگنال گفتاری است که از برخی اشیای اطراف میکروفن به وجود آمده و بعد از چند میلی ثانیه به میکروفون می‌رسد. (انعکاس سیگنال گفتار اصلی در اشیایی که میکروفن را احاطه کرده‌اند.) اگر مکانی که سیگنال گفتار تولید می‌شود، دارای پژواک قوی باشد، پدیده طنین رخ می‌دهد که ممکن است حتی چند ثانیه هم به طول بیانجامد. ابهام

زبان به صورت طبیعی دارای ابهاماتی است. گاهی ما نمی‌توانیم بفهمیم که واقعا چه کلمه‌ای گفته شده است. دو نوع ابهام در بازشناسی گفتار به وجود می‌آید و مخصوص ASR است: تشابه صوتی و ابهام در مرز کلمات. تشابه صوتی، اشاره به کلماتی دارد که صدای یکسان اما املای متفاوتی دارند. در واقع دو کلمه مجزا با صدا و تلفظ یکسان هستند. هنگامی که رشته‌ای از گروههای آوا در قالب رشته‌ای از کلمات گذاشته می‌شود، با ابهام مرز کلمات رو به رو می‌شویم. این ابهام هنگامی پیش می‌آید که چند راه برای گروه‌بندی صداها به کلمات داشته باشیم: مثال زیر این موضوع را به خوبی روشن می‌کند [۱۳].

It’s not easy to wreck a nice beach.

It’s not easy to wreck an ice beach.

سیستم‌های تشخیص‌دهنده گفتار انواع مختلفی دارند، بعضی قادرند گفتار پیوسته را شناسایی نمایند، بعضی دیگر فقط می‌توانند گفتار گسسته (که بین کلمات سکوت وجود دارد) را تشخیص دهند. بعضی سیستم‌ها قادرند کلمات بیان‌شده توسط افراد مختلف را تشخیص دهند و بعضی فقط کلمات یک گوینده را تشخیص می‌دهند. به هر حال ایده آل‌ترین سیستم آن است که بتواند گفتار پیوسته غیر وابسته به گوینده را در محیط نویزی شناسایی نماید.

۲-۵- الگوریتم‌های تشخیص گفتار

از آن‌جایی که عوامل بسیار متفاوتی بر تشخیص گفتار تأثیرگذارند، الگوریتم‌های تشخیص گفتار بسیار پیچیده هستند. این الگوریتم‌ها را می‌توان در سه گروه مبتنی بر الگو، مبتنی بر دانش و آگاهی و مبتنی بر آمار تقسیم‌بندی کرد.

۲-۵-۱- مبتنی بر الگو

در این گروه از الگوریتم‌ها، گفتار ورودی با الگوهای از پیش ضبط‌شده مقایسه می‌شود تا بهترین تطبیق یافت شود. دقت این گروه برای الگوهای موجود خوب است اما به هرحال تعداد الگوها ثابت است و اگر بخواهیم با توجه به شرایط گفته‌شده برای هر کلمه الگوهای متفاوت و متعددی قرار دهیم، اتخاذ این روش به‌طور عملی غیرممکن خواهد شد.

۲-۵-۲- مبتنی دانش و آگاهی

در این الگوریتم‌ها سعی می‌شود مهارت انسان در تشخیص گفتار شبیه‌سازی شود و در سیستم تعبیه گردد. این شیوه اگرچه به نظر بسیار خوب به نظر می‌رسد، اما به دست آوردن این مهارت‌ها و استفاده از آن‌ها در سیستم تشخیص گفتار به‌راحتی میسر نیست و درواقع این روش غیرعملیاتی به‌حساب می‌آید.

۲-۵-۳- مبتنی بر آمار

در این روش‌ها، تغییرات در گفتار به صورت آماری مدل می‌شود و این تغییرات آماری کمک می‌کند تا سیستم تشخیص گفتار امکان یادگیری تدریجی داشته باشد. در سامانه‌های جدید تشخیص گفتار با استفاده از شبکه‌های گسترده عصبی و روش‌های مبتنی بر آمار نتایج بسیار دقیق‌تر و بهتری به دست آمده است. در حال حاضر بسیاری از شرکت‌های مهم مانند IBM و مایکروسافت روی این سامانه‌ها سرمایه‌گذاری کرده‌اند و به نتایج بسیار خوبی رسیده‌اند. یکی از سرویس‌دهندگان تلفن همراه در کشور فرانسه یک پورتال صوتی راه‌اندازی کرده است و اخبار و نتایج مسابقه‌های ورزشی را از این طریق در اختیار مشترکان خود قرار می‌دهد. شرکت ماشین‌سازی هوندا نیز یک سیستم راه نوردی با کمک صوت راه‌اندازی کرده است تا رانندگان بهتر بتوانند خودرو را هدایت کنند. با این پیشرفت‌ها به نظر می‌رسد که در آینده‌ای نه‌چندان دور فناوری تشخیص گفتار بخشی از زندگی و کار هرروزه ما خواهد شد.

این سیستم‌ها با به کار گیری روش‌های مختلف طبقه‌بندی و شناسایی الگو قادر به تشخیص کلمات هستند که البته

برای افزایش دقت در شناسایی از یک فرهنگ لغت نیز در انتهای سیستم استفاده می‌شود. روش‌هایی مانند مدل مخفی مارکوف یا شبکه‌های عصبی در بسیاری از سیستم‌های تشخیص گفتار مورداستفاده قرار می‌گیرند و در بخش‌های انتهایی سیستم از هوش مصنوعی کمک گرفته می‌شود. امروزه با داشتن میکروفن و کارت صوتی در کامپیوتر و به کارگیری نرم‌افزار تشخیص گفتار می‌توان دستورات یا کلمات را به صورت صوتی به کامپیوتر وارد کرد. حتی در بعضی از گوشی‌های تلفن همراه از این سیستم‌ها جهت دریافت دستورات به صورت صوتی استفاده می‌شود.

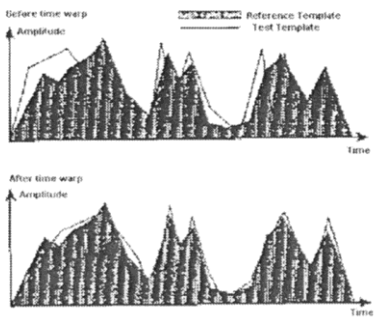
۲-۶-روشهای پیاده سازی موجود

۲-۶-۱- روش پیچش زمانی پویا(DTW)

این روش بر یکسان‌سازی و تطبیق الگوها، استوار است. مبتنی بر الگوست. به دلیل اینکه یکی از خواص رشته، پیچش (کشیدن یا فشرده شدن در زمان) برای اندازه شدن با دیگر رشته‌هاست، برای انجام این کار از برنامه نویسی پویا استفاده می‌شود، این روش تطبیق پیچش پویا (DTW) نام دارد [۱۷].

در این روش مشخصه‌هایی از سیگنال صحبت استخراج می‌شود و حاصل استخراج این مشخصه‌ها بعد از مقایسه با الگوهای از پیش تعریف شده و در نظر گرفتن معیارهای مناسب برای مقایسه، کلمه ادا شده را تعیین خواهد کرد.

قسمت مهم این روش، استفاده از یک تابع بهینه فشرده سازی زمانی است که برای ایجاد صف‌بندی زمانی غیرخطی به کار می‌رود. شکل ۲-۱ الگوها را پیش و پس از اعمال این روش نشان می‌دهد [۱۸].



شکل ۲-۱: نمونه یک الگو پیش و پس از اعمال روش پیچش زمانی پویا

DTW در کاربردهای ساده و الگوریتمهای آسان با حداقل سخت افزار به کار می‌رود. این روش برای بازشناسی کلمات مجزا (IWR) تولید شد اما کاربردهای دیگری نیز دارد مثلا در CSRهایی که از روش گفتار پیوسته استفاده می‌کنند.

چون DTW احتیاج به الگوی قابل دسترسی برای تطبیق هر گفتاری دارد (منبعهای مختلف زیادی در گفتار وجود دارد)، زیاد به کار برده نمی‌شود. عموما برای کارهای پیچیده با واژگان بزرگ و بازشناسی گفتار پیوسته استفاده نمی‌شود.

۲-۶-۲- روش مدل مخفی مارکوف(HMM)

همانطور که گفته شد، روش DTW در سیستمهای مستقل از گوینده با واژگان کوچک (نزدیک ۱۰۰کلمه) و سیستمهای وابسته به گوینده با واژگان متوسط (تقریبا ۵۰۰ کلمه) بد نیست، اما با بزرگ شدن سیستم و با افزایش تعداد الگوهای کافی و هزینه محاسبه و جستجو، استفاده از این روش غیرممکن می‌شود.

اگر امکان به کارگیری DTW با واحدهای زیرکلمه به جای الگوهای مرجع وجود داشته باشد، میزان اطلاعاتی که باید ذخیره شود، کاهش می‌یابد [۱۹] اما در حالت محاسبه، اثرات مرزهای فشرده واحدهای زیر کلمه، مشکل‌آفرین خواهد بود.

در سال ۱۹۷۰ و مخصوصا ۱۹۸۰، محققان گفتار، شروع به یافتن روشهای اتفاقی برای مدل‌سازی گفتار مخصوصا در سیستمهای بزرگ کردند تا مشکل روش قبلی را از بین ببرند.

این روش برای نشان دادن اینکه مدلهای خصوصیات ذاتی، برخی از تغییرات در گفتار را به کار می‌برند، استفاده می‌شود [۱۷]. در واقع دنباله موقتی از نشانه‌های مشاهده شده طیف ایجاد می‌شود که می‌تواند به صورت یک زنجیره مارکوف مدل‌سازی شود تا روشی را که یک صدا به صدای دیگری تبدیل می‌شود، توصیف کند و از بیشترین مقدار رشته احتمالاتی که منجر به تولید همان صدای انسان می شود، برای تطبیق استفاده می‌کند. در این روش، الگوریتم بازشناسی صحبت از روی داده‌های آموزشی مدل در نظر گرفته شده، مشخصات آماری هر کلمه را استخراج می‌کند و یک تابع چگالی احتمال برای مشاهده نسبت به مدل تخمین می‌زند. به عبارت دیگر در این روش یک مدل از

پارامترهای احتمالی برای تولید صوت توسط انسان در نظر گرفته می‌شود و با استفاده از داده‌های آموزشی سعی می‌شود، پارامترهای این مدل طوری تخمین زده شوند که سلسله مشاهدات، دارای بیشترین شباهت به مدل باشند.

دو روش تصادفی وجود دارد HMM و ANN . با توجه به سابقه، HMM بیشتر در پردازش گفتار استفاده شده است. کار روی HMM از ابتدای سال ۱۹۷۰ شروع شد. این روش هم در سیستمهای وابسته به متن و هم مستقل از متن مورد استفاده قرار می‌گیرد.

یکی از فواید این روش، این است که کاملا صفات اختصاصی ساختار صوتی گفتار مدل شده را از بین می‌برد [۱۷]. روش HMM ابزاری موثر برای الگوریتم رمزگشایی و الگوریتم آموزشی با سرپرستی خودکار را فراهم می‌سازد.

HMM پایه دارای مدل‌سازی سطح پایین ضعیف است که موجب سردرگمی بین کلمات مشابه اکوستیکی می‌شود. همچنین مدل‌سازی ضعیف معنایی با فهم سطح بالای گفتار، کاربردها را به موقعیتهایی که تعداد حالت محدود باشد، منحصر می‌کند.

محدودیت دیگر HMM این است که مدل‌سازی مستقیم تلفظ در آن مشکل می‌شود و الگوریتمهای آموزشی HMM نمی‌تواند ساختار توپولوژیکی یا مدلهای کلمه و زیرکلمه را بیاموزد.

۲-۶-۳- روش شبکه عصبی مصنوعی(ANN)

این شبکه‌ها می‌توانند مساله پیچش زمانی پویا در الگوهای ورودی را حل کنند. شبکه‌های عصبی فواصل محلی قاب تا قاب هر یک از الگوهای مرجع را به صورت موازی محاسبه می‌کند (هم برای سیستمهای بازشناسی پیوسته و هم گسسته). در سیستمهای بازشناسی پیوسته این فواصل محلی تابعی از توزیعهای احتمال است.

در بازشناسی گسسته ابتدا فشرده‌سازی برداری انجام می‌شود و هر ورودی با نماد ویژه‌ای برچسب می‌خورد. با استفاده از جداول کشف که احتمالات نماد هر یک از الگوهای ورودی را داراست، می‌توان فواصل محلی را محاسبه نمود.

۲-۷-جمع آوری پایگاه داده و عوامل موثر بر آن

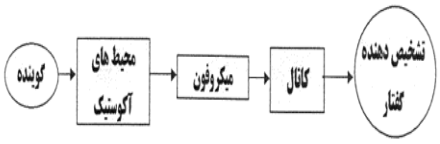
در سیستمهای بازشناسی خودکار گفتار که بر پایه مدلهای آکوستیکی زیر کلمه مثل واج یا هجا شکل می‌گیرند، واژگان سیستم باید توانایی توصیف ورودی‌ها (کلمات یا جملات) را به تمام حالتهای قابل دریافت داشته باشد. هدف محققان تهیه مدلهای تلفظی است که شامل گونه‌های مختلف تلفظ کلمات باشند. وقوع انواع تلفظ ها در گفتار پیوسته یک پدیده شناخته شده آوایی است. تاثیر متقابل کلمات و تغییرات موضعی و مکانیزم های دیگر همگی باعث ایجاد گونه‌های مختلف تلفظ یک کلمه می‌گردند. بعضی از این پدیده‌ها وابسته به گوینده هستند در حالی که برخی دیگر به گوینده وابسته نیستند [۲۰].

وقتی گونه‌های تلفظی کلمات توسط مدلهای تلفظی پوشش داده شوند، صحت بازشناسی افزایش می‌یابد، در واقع در حین بازشناسی، ورودی‌ها بهتر با تلفظ اصلی تطبیق می‌یابند و بنابراین فرآیند بازشناسی تسهیل می‌شود. مدلهای تلفظی در واژگان، به منظور متمایز نمودن ورودی‌ها در سطح کلمات ایجاد می‌شوند، با این حال معرفی گونه‌های تلفظی زیاد در واژگان افزایش سردرگمی سیستم را نیز به دنبال خواهد داشت و این مساله می‌تواند صحت بازشناسی را کاهش دهد [۲۱]. بنابراین باید محدودیتهایی در پذیرش گونه‌ها در نظر گرفته شود [۲۲]. مشاهده می‌شود ایجاد مجموعه واژگان مناسب در این نوع سیستمهای بازشناسی، وابسته به عوامل زیادی از جمله زبان گفتار و آشنایی با قواعد تلفظی و سایر قواعد آن زبان است، به همین دلیل در این پروژه ، ما سعی کرده‌ایم از روشی استفاده کنیم که وابستگی زیادی به زبان گفتار نداشته باشد.

یکی از عوامل دیگری که کارایی سیستم بازشناسی گفتار را کاهش می‌دهد، حضور نویز است. هرگاه گفتاری که باید بازشناسی شود، با نویز تخریب گردد، کارایی سیستم بازشناسی به شدت کاهش می‌یابد. اثر تخریبی نویز بر روی سیگنال گفتار آن است که توزیع بردارهای ویژگی گفتار تخریب شده، شبیه توزیعی که سیستم بازشناسی با آن آموزش دیده، نیست. این عدم تطابق باعث کاهش کارایی سیستمهای بازشناسی در شرایط نویزی می‌شود.

۲-۷-۱- نویز و ضریب اطمینان

شکل ۲-۲ نشان می‌دهد که یک صدا از زمانی که توسط گوینده بیان می‌شود تا به سیستم برسد از چه مراحلی عبور می‌کند [۲۳].



شکل ۲-۲ : مراحل عبور صدا از گوینده تا سیستم

به هنگام ایجاد یک فایل صوتی، عوامل زیادی وجود دارد که امکان تغییرپذیری این صوت را ایجاد کرده و بازشناسی را با خطا مواجه می‌سازند. نویز در واقع اطلاعات و صداهای ناخواسته در سیگنال گفتار است.

مصنوعات غیر گفتار به شدت متنوع هستند:

- نویزهای زمینه شامل نویزهای زمینه ثابت مثل صدای موتور ماشین، خش خش هوا و نویزهای زمینه یا پیش‌نمای متناوب مثل زنگ تلفن یا پارس سگ.
- نویزهای میکروفون و کانال (کلیک، ...) که با تغییر توابع سیستم به وجود می‌آید. مثل کانالهای متفاوت (خط زیر زمینی یا سلولی یا ...) یا میکروفونهای متفاوت.
- نویزهای غیرحرفی گوینده (عطسه، خنده، صدای لب، ...) می‌توانند با گفتار همزمان شوند.

روشهای زیادی برای از بین بردن یا کم کردن تاثیر نویزها بر گفتار وجود دارد و محققان زیادی روی این موضوع کار کرده‌اند ([۲۲] و [۲۳])که از حوصله این بحث خارج است.

۳- پیشینه پژوهش

پژوهش در فناوری صدا و گفتگو پیش از ظهور کامپیوترهای دیجیتال آغاز شده است. این پژوهشها با یک پروژه سنتز گفتار در آزمایشگاه‌های بل در سال ۱۹۳۶ شروع شد که منجر به ابداع دستگاهی به نام The Vodor گشت. دستگاهی که در نمایشگاه جهانی سال ۱۹۳۹ ارائه شد. ارتباط بین گفتار و ریاضیات منجر به دستیابی به اولین موفقیت در اوایل دهه ۱۹۷۰ شد. لئونارد ای بوم و لوید آر ولچ یک راهکار برای شناسایی بر مبنای مفاهیم ایستا با نام مدل پنهان مارکوف اختراع کردند.

در سال ۱۹۶۱ شرکت بل سیستم یک متدولوژی جدید شماره‌گیری تن را توسعه داد. بل از تلفنهای با تکنوژی تن دوگانه- چند فرکانسه(DTMF) استفاده کرد که اولین تلفن از این نسل بود. این تلفن را در نمایشگاه بین المللی سیاتل به نمایش گذاشتند. تلفن های DTMF استفاده از سیگنال‌دهی داخلی را فعال می‌کنند. با وجود افزایش استقرار و استفاده از فن‌آوری‌های تلفن گویا در اوایل ۱۹۷۰ برای انجام وظایف به صورت خودکار در مراکز تلفن، این فناوری همچنان پیچیده و گرانقیمت بود.

همان گونه که مراکز تلفن در اواخر دهه ۱۹۹۰ به سمت چندرسانه‌ای شدن می‌رفتند، شرکتها شروع به سرمایه-گزاری روی کامپیوترهای ادغام شده با تلفن (CTI) با سیستمهای تلفن گویا کردند.

آر جی آبورن در تحقیقی به این نتیجه رسید که با پیشرفت این فناوری، سیستمها می‌توانند از بلندگوهای مستقل شناسایی صدا که واژگان محدودی را پشتیبانی می‌کنند، برای پاسخ به درخواستهای کاربران استفاده کنند. در دهه‌های بعدی فن‌آوری فرامین صوتی بسیار رایج‌تر و ارزانتر در دسترس عموم قرار گرفت به صورتی که در هدفون‌ها و ماشین‌ها نیز از این فناوری استفاده شد.

اندرو هانت و همکارانش در تحقیقی اعلام کردند که دستگاه‌های دریافت کننده فرامین صوتی، دستگاه‌هایی هستند که از طریق فرمانهای صوتی انسانها کار می‌کنند. با حذف نیاز به فشردن کلید و یا استفاده دستی از ابزارهای دستگاه‌ها، مصرف‌کنندگان می‌توانند حتی زمانی که دست آنها به کاری و یا وسیله‌ای مشغول است با استفاده از فرامین صوتی از دستگاه خود استفاده کنند.

در سال ۲۰۰۷ سی ان ان تجاری اعلام کرد که شرکت‌های اپل و گوگل قصد ساخت افزونه‌های تشخیص صدا را دارند. در سالهای بعد شرکت اپل محصول خود را با نام SIRI و شرکت گوگل محصول خود را با نام S-Voice ارائه کردند که این نرم افزارها فرامین صوتی را دریافت و تحلیل می-کردند و سپس با استفاده از پایگاه داده هوشمند، پاسخ مناسب را در اختیار کاربر قرار می‌دادند.

۴-روش پیشنهادی

۴-۱پردازش گفتار

پردازش گفتار را می‌توان به چند حوزه اصلی، تقسیم کرد از جمله [۲۴]:

- **تشخیص گفتار** سیگنال گفتار به جریانی از نمادها (واج و کلمه) ترجمه می‌شود؛ این نمادها، اطلاعات را در سخن گفتن نشان می‌دهند.
- **تشخیص گوینده** تعیین اینکه کدام گوینده از میان مجموعه‌ای از گویندگان ممکن، گفتار داده شده را ادا کرده است.
- **تصدیق گوینده** تایید اینکه آیا گوینده‌ای که سخنی را ادا کرده است، فرد موردنظر است یا نه.
- **سنتز گفتار** تولید مصنوعی گفته‌ای که پیش از این توسط گوینده‌ای ادا نشده است.
- **کدینگ گفتار** تبدیل گفتار به یک نمایش کارآمد (برای اهداف انتقال و یا ذخیره سازی) که امکان بازسازی گفتار اصلی را فراهم می‌کند.

این پژوهش عمدتا بر تشخیص گفتار متمرکز است اما تکنیکهایی که مورد بحث قرار می‌گیرد، کاربردهایی در همه زمینه‌های اصلی دارند.

۴-۲-تشخیص فعالیت صوتی

تشخیص فعالیت صوتی (VAD)، که تحت عنوان تشخیص فعالیت گفتار یا تشخیص گفتار نیز شناخته می‌شود، تکنیکی است که در آن حضور یا عدم حضور گفتار انسان شناسایی می‌شود. کاربردهای اصلی VAD در کد کردن و شناسایی گفتار است. این تکنیک می‌تواند پردازش گفتار را تسهیل کند و نیز می‌تواند برای غیر فعال کردن بعضی پردازشها در طی یک بخش غیر گفتاری یک جلسه صوتی مورد استفاده قرار بگیرد.

مجموعه‌ای از سیگنالهای صوتی در قالب فریم قرار می‌گیرند. در هر لحظه میانگین مقادیر سیگنالها در هر فریم از رابطه زیر محاسبه می‌شود.

۴-۳-استخراج ویژگی

یکی از مرحل اصلی سیستم های تشخیص گفتار استخراج ویژگی می باشد که در این مرحله، سیگنال گفتار به دنباله‌ای گسسته از بردارهای ویژگی تبدیل می‌شود؛ این بردارهای ویژگی تنها شامل اطلاعاتی است که برای تشخیص صحیح گفتار مهم است. استخراج ویژگی به منظور کاهش اندازه داده‌های گفتار و پردازشهای دیگر برای انطباق این داده‌ها با نیازهای طبقه بندی کننده است. استخراج ویژگی استاندارد شامل مراحل زیر است:

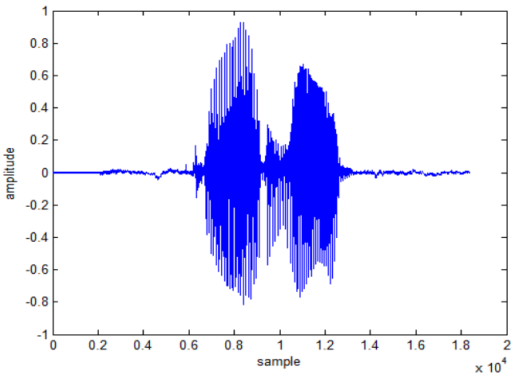
- **بخش‌بندی** سیگنال گفتار به بخش‌هایی تقسیم می‌شود که می‌توان شکل موج را در آنها ثابت در نظر گرفت (مدت زمان معمول ۲۵ میلی ثانیه). طبقه بندی‌کننده‌ها معمولا فرض می‌کنند که ورودی‌شان دنباله‌ای از بردارهای پارامتر گسسته است که در آن هر بردار پارامتر یکی از چنین بخش - فریم‌هایی را نشان می‌دهد.
- **اسپکتروم** روشهای فعلی استخراج ویژگی عمدتا بر اساس طیف فوریه کوتاه مدت و تغییرات آن در زمان هستند، بنابراین طیف فوریه توان و دامنه برای هر بخش گفتار محاسبه می‌شود.
- **تغییرات شنوایی مانند** تغییرات ملهم از یافته‌های فیزیولوژیکی و روانشناختی در مورد درک انسان از بلندی صدا و حساسیتهای مختلف برای فرکانسهای مختلف روی طیف هر فریم گفتار انجام می‌شود.
- **Decorrelation** برخی از تکنیکهای decorrelation برداری برای انطباق بهتر ویژگی‌ها با نیازهای طبقه‌بندی کننده استفاده می‌شود.
- **مشتقات** بردارهای ویژگی معمولا با مشتقات مرتبه اول و دوم خط سیر زمانی آنها (ضرایب دلتا و شتاب) تکمیل می‌شوند. این ضرایب تغییرات و سرعت تغییرات ضرایب بردار ویژگی را در زمان توصیف می‌کنند.

ضرایب کپسترال فرکانس مل

MFCC بر ادراک شنوایی انسان استوار است که نمی‌تواند فرکانسهای بیش از یک کیلوهرتز را درک کند [۲۸]. ویژگی‌های به دست آمده از الگوریتم MFCC ، به پهنای باند فرکانسی شنوایی انسان شبیه هستند. MFCCها برای مدت زمان قابل توجهی ویژگی‌های برتر برای شناسایی گفتار بوده‌اند. موفقیت این ضرایب، به دلیل توانایی‌شان

برای نمایش طیف دامنه گفتار به یک فرم فشرده بوده - است.

قبل از استخراج ویژگی، نمونه صدا باید تحت تبدیل آنالوگ به دیجیتال و پس از آن، پیش تاکید و فیلترینگ قرار بگیرد. برای جلوگیری از aliasing ، نرخ نمونه برداری باید کافی باشد. با توجه به قضیه نمونه‌برداری Nyquists ، حداقل فرکانس نمونه برداری از یک سیگنال با حداکثر فرکانس f باید 2f هرتز باشد. نمونه صدای دیجیتال شده برای کلمه "جلو" در شکل ۴-۱ نشان داده شده است.



شکل ۴-۱ : نمونه صدای دیجیتال شده برای کلمه جلو

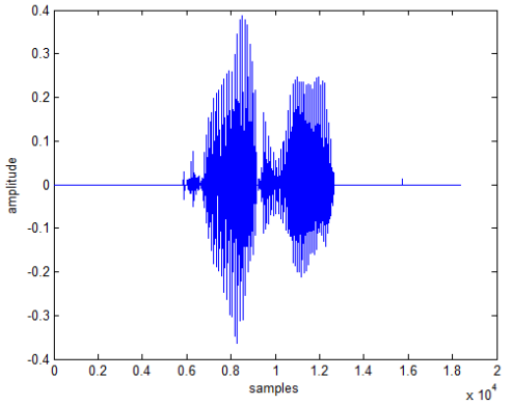
مرحله پیش تاکید، دامنه فرکانس بالاتر را با توجه به فرکانسهای پایین‌تر افزایش می‌دهد. فیلتر FIR مزبور و خروجی گسسته متناظر با آن، به ترتیب در معادلات۴-۱ و ۴-۲داده شده است.

(۴-۱)

(۴-۲)

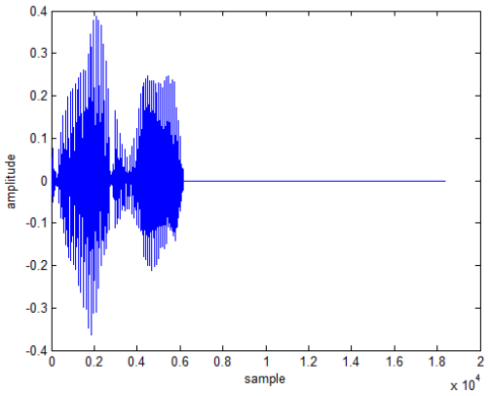
y خروجی است و s ورودی فیلتر FIR است.

سپس گیت نویز بر نمونه صدای پیش تاکید شده اعمال می‌شود. این گیت، دامنه‌هایی (نویز) را که زیر یک مقدار آستانه مشخص قرار می‌گیرند، حذف می‌کند. برآورد طیفی کلمه "جلو" پس از پیش‌پردازش و گیت نویز در شکل ۴-۲ نشان داده شده است.



شکل ۴-۲ برآورد طیفی کلمه جلو پس از پیش‌پردازش و عبور از گیت نویز

پس از این که گیت نویز اعمال شد، نمونه صدا روی محور زمان برای شروع از صفر تراز می‌شود. این کار تراز صفر نامیده می‌شود. بعدا در برنامه، این کار باعث کاهش حجم بار برای فرآیند تطبیق الگو می‌شود. زیرا به این ترتیب نمونه‌های صدا بیش از حالت اولیه به یکدیگر نزدیک خواهند بود. تمام عملیات سیگنالی فوق قبل از استخراج MFCC انجام می‌شوند. عملیات مزبور از تداخل نویز با ویژگی‌های مهم جلوگیری می‌کنند. تراز صفر طیف کلمه "جلو" در شکل ۴-۳ نشان داده شده است.



شکل ۴-۳ تراز صفر برآورد طیفی کلمه جلو

نمونه صدا یک سیگنال متغیر با زمان است. از این رو، باید به صورت فریمهایی با طولی در محدوده ۲۰ تا ۳۰ میلی ثانیه فریم‌بندی شود. برای این که بتوانیم یک برآورد طیفی مطمئن برای هر فریم به دست آوریم، طول فریم نباید بیش از حد کوتاه باشد. از سوی دیگر، نباید بیش از حد طولانی باشد تا تحت یک فریم خاص، نمونه صدا مستقل از زمان باشد (با زمان تغییر نکند). فریمهای مجاور با M از هم جدا شده‌اند(M <N). در این پژوهش همپوشانی فریمها ۱۰۰ و طول فریم ۲۵۶ است. سپس هر فریم با

پنجره همینگ ضرب می‌شود. تابع پنجره همینگ در ۴-۳ بیان شده است. خروجی هر فریم پس از فیلترکردن، از رابطه۴-۴ و به دست می‌آید.

$$W[n] = 0.54 - 0.46\cos\left[\frac{2\pi n}{N-1}\right] \quad (۴-۳)$$

$$Y[n] = X[n] \times W[n] \quad (۴-۴)$$

N تعداد نمونه‌ها در هر فریم، Y سیگنال خروجی، X سیگنال ورودی و W، n امین ضریب پنجره همینگ است.

تبدیل فوریه سریع (FFT) بر هر فریم اعمال می‌شود و سیگنال را به حوزه فرکانس انتقال می‌دهد. ما معمولا یک FFT ۵۱۲ نقطه‌ای انجام می‌دهیم و تنها ۲۵۷ ضریب اول را نگه می‌داریم. به این ترتیب طیف هر فریم به دست می‌آید. اما، این طیف هنوز هم شامل اطلاعات زیادی است که برای مرحله تطبیق ویژگی لازم نیست. الگوریتم تطبیق ویژگی نمیتواند تفاوت بین دو فرکانس با فاصله نزدیک را تشخیص دهد. به همین دلیل دسته‌ای ازبین‌های طیفی را می‌گیریم و مجموع آنها را به دست می‌آوریم تا بفهمیم چه مقدار انرژی در مناطق فرکانسی گوناگون وجود دارد.

اولین فیلتر بسیار باریک است و نشان می‌دهد چه مقدار انرژی نزدیک صفر هرتز وجود دارد. هر چه فرکانس بیشتر می‌شود، فیلترهای ما پهنتر می‌شوند؛ زیرا تغییرات کمتر اهمیت دارد. معادله محاسبه MEL برای یک فرکانس داده شده در رابطه ۴-۵ نشان داده شده است.

(۴-۵)

$$F(MEL) = 2595 \times \log_{10}[1 + f / 700]$$

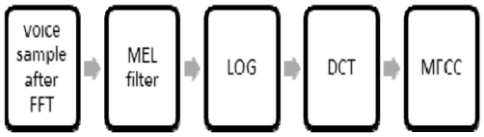
ما تنها به مقدار تقریبی انرژی در هر نقطه علاقه مند هستیم. در اینجا مجموعه‌ای از ۲۶ فیلتر مثلثی اتخاذ شده است. برای محاسبه انرژی‌های بانک فیلتر، هر بانک فیلتر را با طیف انرژی ضرب می‌کنیم و سپس ضرایب را جمع می‌نماییم. هنگامی که این کار انجام شد، ۲۶ عدد خواهیم داشت که به ما نشان می‌دهد چه مقدار انرژی در هر بانک فیلتر وجود دارد. لگاریتم این ۲۶ مقدار انرژی گرفته می‌شود و در ادامه تبدیل کسینوسی گسسته (DCT) انجام خواهد شد. DCT با استفاده از معادله ۴-۶ محاسبه می‌شود.

$$C_n = \sum_{k=1}^{k} (\log S_k)\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{K}\right] \quad (۴-۶)$$

$$n = 1,2,...,K$$

$$S_k = FFTcoefficients$$

مقدار K ، ۲۶ در نظر گرفته شده است. بنابراین ما با ۲۶ ضریب مواجهیم، اما برای تطبیق ویژگی؛ تنها ۱۲-۱۳ ضریب پایینتر نگه داشته می‌شوند. ویژگی‌های حاصل (۱۲ عدد برای هر فریم) ضرایب سپسترال فرکانس مل نامیده می‌شوند. بنابراین نمونه‌ای که پس از اعمال FFT در حوزه فرکانس است با استفاده از فیلتر MEL و DCT به حوزه زمان بازگردانده می‌شود؛ همانطور که در شکل۴-۴ نشان داده شده است.

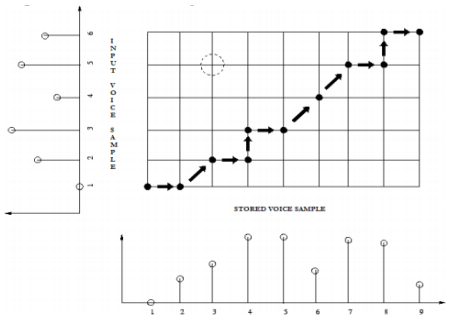


شکل ۴-۴ : مراحل تبدیل از حوزه فرکانس به حوزه زمان

۴-۴-تطبیق ویژگی (DTW)

برای مواجهه با سرعتهای متفاوت در صحبت کردن، در تشخیص گفتار، از انحراف زمانی پویا (DTW) استفاده می‌شود. DTW الگوریتمی است که برای اندازه گیری شباهت دو دنباله، که در زمان و یا سرعت متفاوتند، مورد استفاده قرار می‌گیرد.

در این مرحله، ویژگی‌های کلمه که در گام قبلی محاسبه شده با قالبهای مرجع مقایسه می‌شود. برای محاسبه حداقل فاصله بین ویژگی‌های کلمه ادا شده و قالبهای مرجع، الگوریتم DTW اجرا می‌شود. کمترین مقدار در میان نمرات محاسبه شده جواب مطلوب را به دست می‌دهد. اگر یک سری زمانی به صورت غیر خطی با کشیده شدن و یا جمع شدن در امتداد محور زمان "منحرف" شده باشد، DTW قادر است تراز بهینه را بین این سری زمانی و یک سری زمانی دیگر پیدا کند؛ میزان تطبیق بین دو سری زمانی با فاکتور فاصله اندازه‌گیری می‌شود. انحراف زمانی پویا برای دو نمونه صدا در شکل ۴-۵ نشان داده شده است.



شکل ۴-۵ : Dynamic time wrapping برای دو نمونه صدا

یک ماتریس از مرتبه n در m ایجاد می‌شود که عناصر (i,j) آن، فاصله $d(a_i,b_j)$ بین نقاط a_i و b_j در دو سری زمانی است. محاسبات اقلیدسی برای اندازه گیری فاصله بین ویژگی‌های نمونه ورودی و قالب ذخیره شده استفاده می‌شود. سپس فاصله با رابطه ۳-۱۴ اندازه گیری می‌شود.

(۷-۴)

$$D(i,j)=\min\left[D(i-1,j-1),D(i-1,j),D(i,j-1)\right]+d(i,j)$$

قالب مربوط به کمترین فاصله، کلمه شناسایی شده است.

دو ادای یک کلمه خاص، توسط یک کاربر، هم می‌توانند از نظر زمانی با هم متفاوت باشند. برای مثال، two می‌تواند به صورت to یا too تلفظ شود. DTW با تراز کردن مناسب کلمات و محاسبه حداقل فاصله بین دو کلمه این مساله را حل می‌کند.

۵-تحلیل نتایج

۵-۱-مرحله آموزش

همیار فرد نابینا از این طریق اماکن مورد نظر را بر روی نقشه مشخص می‌کند و عنوان هر مکان از طریق میکرفون توسط شخص نابینا چندین بار قرائت می‌شود. از این طریق موقعیت جغرافیایی اماکن موردنظر به همراه عنوان صوتی آنها در پایگاه داده ذخیره می‌شود.

۵-۲-مرحله تست

برای تست نرم افزار در هر لحظه، موقعیت جغرافیایی از PositionSource دریافت می‌شود که اگر شخص نابینا به GPS مجهز باشد، هر لحظه این اطلاعات به صورت واقعی دریافت می‌شود و بر روی نقشه نگاشت می‌گردد. در این پروژه مختصات جغرافیایی به صورت شبیه سازی شده استفاده می‌شود؛ بدین ترتیب که کاربر از طریق مشخص کردن مختصات جغرافیایی خود به روی نقشه، محل کنونی خود را مشخص کرده و سپس از طریق کلیدهای جهتی صفحه کلید ناوبری بر روی نقشه را به صورت دستی انجام می‌دهد.

زمانی که شخص نابینا از طریق میکرفون عنوان مکانهای از پیش تعریف شده را قرائت می‌کند، ابتدا از طریق VAD صوت بخش‌بندی شده و در مرحله بعدی از طریق روش MFCC ویژگی استخراج می‌شود و در ادامه ویژگی استخراج شده با کل ویژگی‌های ذخیره شده در پایگاه داده به صورت brute force و با الگوریتم DTW مقایسه می‌شود. ویژگی‌ای که دارای کمترین فاصله (بیشترین امتیاز) باشد، انتخاب می‌شود و مختصات جغرافیایی منتسب به این ویژگی از پایگاه داده استخراج می‌شود. حال از سرویس GeoRoute برای مسیریابی استفاده می‌گردد. سرویس GeoRoute اطلاعات سطح در مورد یک مسیر همانند طول مسیر، زمان تخمینی جهت طی مسیر و اطلاعات کافی جهت رندر تصاویر را در اختیارمان قرار می‌دهد.

اطلاعات مسیرها در قالب GeoRouteSegment است که اطلاعات هر بخش را با جزئیات بیشتری بیان می‌کند. سرویس GeoRoute جهت مسیریابی، مختصات جغرافیایی مبدا و مقصد را دریافت می‌کند، به همین ترتیب زمانی که شخص نابینا عنوان مکانی را قرائت می‌کند، پس از جستجو، نقطه مورد نظر به عنوان مقصد و نقطه جاری به عنوان مبدا در نظر گرفته می‌شود. سپس از سرویس‌دهنده اطلاع مسیر مورد نظر استخراج می‌گردد. بدین ترتیب اطلاعات جهت حرکت به صورت فایل صوتی در بازه مشخص برای شخص نابینا پخش می‌شود و از این طریق شخص نابینا مسیریابی را انجام می‌دهد.

۶-نتیجه گیری

سیستم همیار نابینایان از دو بخش شناسایی گفتار و ناوبری بر روی نقشه تشکیل شده است. دربخش گفتار سیگنالهای ورودی از طریق الگوریتم ADT بخش‌بندی شده و سپس ویژگی MFCC از آنها استخراج می‌شود. ویژگی‌های حاصل، از طریق روش تطبیق DTW با هم مقایسه می‌شوند. سیستم علاوه بر ۴ جهت اصلی، با

ذخیره‌سازی ویژگی‌های عناوین اماکن مختلف (حداکثر ۵ نام متفاوت) نیز آموزش داده شده است. نتایج نشان داده که به ازای هر کلمه به ۱۰ ویژگی برای به دست آوردن بهترین نتیجه نیاز است. در واقع این تعداد، موازنه‌ای بین سرعت و دقت برقرار می‌کند. کارایی عملیات جداسازی مطلوب و در حدود ۹۰ درصد است. فاصله مابین کلمات یکسان کمتر از ۱۰۰ و کلمات متفاوت بیشتر از ۳۰۰ است. بنابرین حد آستانه ۱۵۰ مقدار مناسب جهت تمایز کلمات موجود در پایگاه داده از سایر کلمات است. در بخش ناوبری بر روی نقشه، از پایگاه داده، موقعیت جغرافیایی مکان مورد نظر بازیابی شده و با توجه به نقطه فعلی شخص نابینا، مسیر بخش‌بندی می‌شود. سپس، سیستم، هر بخش را از طریق فایل صوتی مرتبط قرائت می‌کند. همچین در سیستم یک شبیه‌ساز GPS پیاده‌سازی شده که از طریق آن می‌توان سیستم را به صورت مجازی تست کرد.

۶-۱-کارهای آینده

پیدا کردن روشهای تشخیص گفتار خودکار که برای واژه های فارسی کارآمد باشد، توجه بسیاری را به خود معطوف داشته زیرا تحقیقات در این زمینه محدود باقی مانده است. در این کار، استحکام MFCC در ترکیب با الگوریتم DTW به وضوح نشان داده شد. علاوه بر این، تکنیک آشکارساز فعالیت صوتی نیز تاثیر قابل توجهی بر عملکرد سیستم دارد. برای بهبود نتایج می‌توان نقطه شروع و نقطه پایانی را با استفاده از مفهوم ترکیبی یافته‌های انرژی و نرخ عبور از صفر به طور دقیق تعیین کرد. زیرا یافته‌های انرژی نویز را حذف می‌کند و پریود سکوت همچنان در سیگنال حاضر است و عبور از صفر برای آشکارسازی صداهای ضعیف به کار می‌رود. بنابراین، ثابت شده که ترکیب مفاهیم بالا و استفاده از DTW و همبستگی برای پیدا کردن بهترین انطباق، یک روش موثر در تشخیص گفتار است که نتایج قابل توجه و بهتری تولید می‌کند. از سوی دیگر، اضافه کردن ضرایب دلتا و دلتا دلتا در بهبود دقت تشخیص کلی کمک می‌کند که در آینده می‌تواند مورد بررسی قرار گیرد. باید آزمایشات بسیاری انجام شود تا بهترین پارامترهایی انتخاب شوند که کارایی تشخیص گفتار فارسی را به حداکثر می‌رسانند. علاوه بر این، برای فرهنگ لغات بزرگتر می‌توان از روشهای دیگری نظیر HMM و رویکرد مبتنی بر شبکه عصبی مصنوعی استفاده کرد اما روش ارائه شده برای مجموعه فرهنگ لغات کوچک، کارآمد و نسبت به روشهای دیگر مقرون به صرفه است.

فهرست مراجع

[۱] H. M. Kamel, "A Rapid Internet Browsing Technique for Visually Impaired Web Users," *Towards Accessible Search Systems,* pp. 36-41, 2010.

[۲] *WHO | Prevention of avoidable blindness and visual impairment*. Available: http://www.who.int/blindness/en/

[۳] ۲۰۱۰*World Population Data Sheet - Population Reference Bureau*. Available: http://www.prb.org/Publications/Datasheets/2010/2010wpds.aspx

[۴] N. Rajput, S. Agarwal, A. Kumar, and A. Nanavati, "An alternative information web for visually impaired users in developing countries," *Proceedings of the 10th international ACM SIGACCESS conference on Computers and accessibility,* pp. 289-290, 2008.

[۵] C. Ghaoui, *Encyclopedia of human computer interaction*: IGI Global, 2006.

[۶] M. Rotard, S. Knodler, and T. Ertl, "A tactile web browser for the visually disabled," presented at the Proceedings of the sixteenth ACM conference on Hypertext and hypermedia, 2005.

[۷] S. Meers and K. Ward, "Head-tracking haptic computer interface for the blind," *Faculty of Informatics-Papers,* pp. 746-758, 2010.

[۸] S. Kawanaka, Y. Borodin, J.P.Bigham, D.Lunn, H.Takagi, and C. Asakawa, "Accessibility commons: a metadata infrastructure for web accessibility," in *10th international ACM SIGACCESS conference on Computers and accessibility*, 2008 ,pp. 153-160.

[۹] Z. Wang, X. Xu, and B. Li, "Bayesian tactile face," *Computer Vision and Pattern Recognition,* pp. 1-8, 2008.

[۱۰] L. Spalteholz, K. F. Li, N. Livingsto, and F. Hamidi, "Keysurf: a character controlled browser for people with physical disabilities," in *17th international conference on World Wide Web*, 2008, pp. 31-40.

[۱۱] M. Y. Ivory, S. Yu, and K. Gronemyer, "Search result exploration: a preliminary study of blind and sighted users' decision making and performance," *CHI'04 extended abstracts on Human factors in computing systems, ,* pp. 1453-1456, 2004.

[۱۲] J.H.M.D.Jurafsky, *Speech and Language Processing, An introduction to Natural Language Processing, Computational Linguistics , and Speech Recognition*: Prentice Hall, 2000.

[۱۳] M.Forsberg, "Why is Speech Recognition Difficult?," Department of Computing Science, Chalmers University of Technology, 2003.

[۱۴] E.A.J.Allwood, "Corpus-based research on spoken language," 2001.

[۱۵] J.Holmes, *An introduction to sociolinguistics*: Longman Group UK Limited, 1992.

[۱۶] O. Deroo, *A Short Introduction to Speech Recognition*.

[۱۷] J.R.Deller, J. H. L. Hansen, and J.G.proakis, "Discrete-time Processing of Speech Signals," *IEEE Signal Processing Society,* 1993.

[۱۸] S.J.Young, *The HTK book*: Cambridge Univercity Eng, 2001.

[۱۹] L.R.Bahl, S.K.Das, and P.V.Desousa, "Some Experiments with Large Vocabulary isolated word Sentence Recognition," in *Proceedings of the IEEE International Conference on Acoustics*.

[۲۰] H.Strik and C.Cucchiarini, "Modeling Pronunciation Variation for ASR:A Survey of the Literaure," *Speech Communication,* pp. 225-246.

[۲۱] B.Gold and N.Mogan, *Speech & Audio Signal Processing*: John wiley & sons, 2000.

[۲۲] R.M.Stem, "Robust Signal Representations For Automatic Speech Recognition," Department of Electrical and Computer Engineering and School of Computer Science, Carnegie Mellon University

Institute for Mathematics and its Applications, University of Minnesota, 2000.

[۲۳] I. Kiss, "Noise Robust Speech Recognition," in *Advanced Topics in Signal Processing: Voice and Multimodal Technologies and Services.* , ed.

[۲۴] l. burget, "Complementarity of Speech Recognition Systems and System Combination," Department of Computer Graphics and Multimedia, Brno University of Technology Faculty of Information Technology.

[۲۵] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustic Society of America,* vol. 87, pp. 1738-1752, 1990.

[۲۶] F.Jelinek, *Statistical Methods for Speech Recognition*, 1998.

[۲۷] L. Rabiner and B.H.juang, *Fundamentals of speech recognition Signal Processing*: Prentice Hall, 1993.

[۲۸] B. j. mohan and R. Badu, "Speech Recognition using MFCC and DTW," presented at the ICAEE, 2014.