
بسم الله الرحمن الرحيم

دانشگاه صنعتی امیرکبیر

دانشکده مهندسی برق

پایان نامه کارشناسی ارشد گرایش الکترونیک

مقاوم سازی بازشناسی گفتار با اعمال پردازش زیرباندی

نگارش:

حجت اله یگانه

استاد راهنما:

دکتر سید محمد احدی

بهمن ماه ۱۳۸۷

تقدیم به

پدری مهربان،

مادری فداکار

و همسری دل‌آرام

برخود لازم می‌دانم از زحمات بی شائبه استاد ارجمندم جناب آقای دکتر احدی تشکر و قدردانی نمایم. بی شک، بدون حمایت‌های دلسوزانه خانواده عزیزم پیمودن راه تاکنون امکان‌پذیر نمی‌بود. امیدوارم بتوانم قدردان زحمات ایشان باشم.

این پروژه تحت قرارداد پژوهشی شماره ۵۰۰/۸۱۹۹/ت مورخ ۸۷/۵/۲۸ از حمایت های مـالی و معنوی مرکز تحقیقات مخابرات ایران بهره‌مند شده است.

چکیده

در این پروژه مقاوم سازی بازشناسی گفتار در محیط های نویزی بر مبنای پردازش زیربندی بررسی شده است. مقاوم سازی بازشناسی گفتار یکی از مسائل مهم در این حوزه می‌باشد که کار بر روی آن همچنان ادامه دارد. از روش‌های گوناگونی به منظور تحقق یافتن این مهم استفاده می‌شود و ایده‌های متنوعی نیز در مقالات و تحقیقات ارائه می‌گردند. عیب عمده اکثر روش‌های پیشنهاد شده پیچیدگی زیاد و سرعت کم الگوریتم‌های آن است. ما در این رساله تلاش نموده‌ایم به ارائه روش‌هایی بپردازیم که عیوب مذکور را در حد بسیار کمتری داشته باشند و در عین حال هدف ما را که همانا افزایش دقت بازشناسی گفتار در محیط‌های نویزی است محقق سازند.

از این رو در ابتدا با بررسی روند استخراج ویژگی های مطرح MFCC سعی نمودیم عیب این بردار ویژگی را برطرف نماییم. دلیل این امر آن است که بردار ویژگی MFCC دارای الگوریتم بسیار ساده و سریعی است و در محیط‌های عاری از نویز یا به اصطلاح تمیز از دقت بازشناسی خیلی خوبی برخوردار است. بنابراین اگر بتوان الگوریتمی پیشنهاد نمود که ساختاری شبیه MFCC داشته باشد و در عین حال بردار ویژگی حاصل نسبت به نویز محیط مقاوم تر نشان دهد، این روش از امتیاز بالایی برخوردار می‌گردد. با دنبال کردن روند استخراج ویژگی MFCC به این مسئله پی می‌بریم که برای به دست آمدن هرکدام از مولفه‌های این بردار ، تمام طیف یک فریم در محاسبه تاثیر دارد. این بدان معنی است که آلوده بودن نواحی خاصی از طیف به تمام بردار MFCC سرایت می‌کند و کارآیی این بردار ویژگی را به شدت پایین می‌آورد.

روند کلی ایده‌های پیشنهادی ما حول دو گام اصلی قابل بیان است. در ابتدا سعی برآن داریم تا با فیلتر کردن سیگنال گفتار در حوزه زیربندی میزان تاثیر نویز را کم کنیم. در ادامه و در گام دوم با اعمال وزن بر خروجی زیربندهای حوزه مل میزان مشارکت زیربندهای با کیفیت تر را در به دست آوردن ویژگی‌های پیشنهادی خود بیشتر نموده و از تاثیر زیربندهای با کیفیت کمتر می‌کاهیم.

نتایج آزمایش‌های ما نشان دادند که تا حد خیلی خوبی به اهداف اصلی خود رسیده‌ایم. الگوریتم-های پیشنهادی ما در عین حالی که ساده هستند در محیط‌های نویزی بسیار مقاوم می‌باشند. مقایسه روش‌های پیشنهادی با روش‌های مطرح دیگر بر این مطلب صحه می‌گذارد. روش‌های ارائه شده منجر به بهبود ۳۲ درصدی روش پایه شده است.

واژگان کلیدی: مقاوم سازی بازشناسی گفتار، فیلترکردن در زیرباند، وزن دهی زیربندهای مل

فهرست مطالب

فهرست شکل‌ها

فهرست جدول‌ها

فصل اول

مقدمه

فصل اول

مقدمه

امروزه سیستم های بازشناسی گفتار کاربردهای زیادی پیدا کرده‌اند. در این میان سهم کشورهای پیشرفته از جنبه تکنولوژی به مراتب بیشتر از سایر کشورها بوده است. گستردگی دنیای ماشینی و رایانه هایی که اطراف ما را پوشانده اند محققین را به سمت استفاده از این ابزار سوق داده است. بشر امروز سعی دارد تا با حداقل زمان ممکن بیشترین کارآیی روزانه را داشته باشد. یکی از مشکلات پیش روی این مهم رابطه میان انسان و ابزار مورد نیاز اوست. سیستم های بازشناس گفتار یکی از راه حل هایی است که توجه محققین را به خود معطوف داشته است. از ساده ترین کاربردهای این سیستم می‌توان به تایپ کردن در محیط‌های نوشتاری رایانه‌ای با بیان متن مورد نظر و بدون تایپ دستی اشاره نمود. از طرف دیگرانسان امروز تلاش می‌کند به دلایل اقتصادی تا حد ممکن حضور خود را در محیط های کاری کم کند. از اینرو سیستم های بازشناسی گفتار در مراکز ارائه دهنده اطلاعات تلفنی نیز گسترش یافته‌اند. این گونه مراکز به سیستم‌های رایانه‌ای مجهز شده‌اند که در این رایانه ها سیستم‌های بازشناس گفتار تعبیه نموده‌اند. با تماس با این مراکز می‌توان اطلاعات مورد نیاز خودرا از این رایانه ها دریافت کرد به گونه‌ای که بدون حضور فرد خاصی این سیستم بازشناسی گفتار است که پیام پرسنده را تشخیص می‌دهد و پس از تشخیص آن، داده‌های مورد نیاز را در اختیار طرف مقابل می‌گذارد. کاربردهای این‌گونه

سیستم‌ها به این موارد ختم نمی‌شوند و امروزه همچنان محققین به دنبال زمینه‌های کاری این رشته می-باشند.

اگر بخواهیم نمایی کلی از یک سیستم بازشناسی گفتار ارائه دهیم باید از معرفی نوع سیگنال گفتار آغاز کنیم. همان گونه که می‌دانیم در سیگنال های تصادفی مشخصات آماری مختلفی وجود دارد که با تحلیل آنها می‌توان به اطلاعات مهمی در مورد یک سیگنال دست یافت. اگر این مشخصات آماری مانند میانگین و واریانس داده ها در طول زمان ثابت باشند این سیگنال را ایستان[1] می‌گویند. اما اگر چنین نباشد و این اطلاعات در طول زمان متغیر باشد سیگنال را ناایستان[2] می‌نامند. معمولاً تحلیل سیگنال های ناایستان مشکلات مخصوص به خود را دارد. کافی است به عنوان مثال به این مطلب اشاره نماییم که تمام سواد حوزه فرکانسی ما و تبدیل فوریه و ابزارهای دیگر برای سیگنال های ایستان تعریف شده اند. یکی از راه حل های این مشکل بریدن سیگنال به قطعاتی است که مشخصات آماری سیگنال در آن ها تقریباً ثابت باشد. اصطلاحاً به این قطعات فریم گفته می‌شود. در برخی از کاربردها مانند تجزیه و تحلیل سیگنال های راداری طول این فریم ها متفاوت است. به بیان دیگر بررسی می‌کنند که سیگنال را در چه بازه ای می‌توان ایستان در نظر گرفت و بنابر ماهیت این گونه سیگنال ها بازه ایستان بودن متغیر است. سیگنال گفتار نیز یک سیگنال غیر ایستان است و ازاینرو برای تحلیل آن نیاز به فریم بندی کردن آن داریم. معمولاً طول فریم ها را در مورد سیگنال گفتار ثابت در نظر می‌گیرند و گفتار را در آن بازه ها پردازش می‌کنند. با این فرض که سیگنال گفتار درفریم ها ایستان است می‌توان این فریم ها را به حوزه فرکانس منتقل کرد و در آن حوزه نیز تغییراتی را به سیگنال اعمال کرد. در ادامه، مسئله‌ای که وجود دارد حجم زیاد داده ها و اطلاعات در هر فریم است. برای مثال اگر یک سیگنال گفتار با فرکانس ۸ کیلوهرتز نمونه برداری شده باشد و طول فریم ها را ۲۰ میلی ثانیه در نظر بگیریم –که بازه معمولی هم هست– در آن صورت برای هر فریم مقدار۱۶۰ نمونه را در اختیار داریم. کافی است حجم اطلاعات را برای یک

[1] Stationary
[2] Non Stationary

گفتار ۱۰ دقیقه ای محاسبه کنیم تا به ابعاد مشکل پی ببریم. در این حالت راه چاره ای که دانشمندان جسته اند، مشکل را تا حد زیادی برطرف نموده است و این راهکار، استخراج بردار ویژگی از فریم ها است.

محققین را ه ها و الگوریتم هایی را ارائه نموده اند که یک سری نماینده تحت عنوان ویژگی از هر فریم گفتار استخراج شود به گونه ای که این ویژگی‌ها نماینده ای برای مشخصات و خصوصیات یک فریم گفتار باشد. بردارهای ویژگی متعددی در این راستا معرفی شده اند که از مهمترین آنها در بحث سیگنال گفتار می‌توان به بردار ضرایب کپسترال حوزه فرکانس مل یا MFCC [1]اشاره نمود.

معمولاً هدف از بازشناسی گفتار تشخیص این مطلب است که گفتار مورد نظر به اصطلاح چه می گوید. به عبارت دیگر می‌توان گفت هدف از بازشناسی گفتار تشخیص مطلب ادا شده در این گفتار است. اگر بتوان این مسئله را تشخیص داد آنوقت این گفتار می‌تواند به متن تبدیل شود یا به یک دستور برای راه اندازی یک سیستم الکتریکی یا کاربردهای دیگری که قبلاً اشاره شد. هنگامی که صحبت از تشخیص مطلب یا اطلاعاتی از یک سیگنال آن هم توسط رایانه باشد، چگونگی تبدیل این تشخیص به الگوریتم های ریاضی و به تبع آن رایانه ای سوالی است که باید جواب داده شود. ساختار پیچیده مغز انسان برای تشخیص و پردازش اطلاعات به مرور زمان آموزش دیده است. به این خاطر است که انسان حاوی چنین قدرت بالای تشخیص و تصمیم گیری می‌باشد. اگر این انتظار را از رایانه نیز داریم باید شبیه اتفاقی که در ساختار مغز انسان می افتد را شبیه سازی نماییم.

انسان برای تشخیص یک مطلب سعی می‌کند ماهیت داده جدیدی که می‌بیند را با اطلاعات گذشته مقایسه نماید و بعد از این مقایسه تصمیم گیری کند. پس اگر قرار است الگوریتم های طبقه بندی کننده[2] نیز این عمل را انجام دهند باید در ابتدا با یک سری از داده ها و اطلاعات آموزش ببینند تا ساختار آن ها را با داده های جدید مقایسه کنند و آنگاه تشخیص دهند که ماهیت و اطلاعات یک ورودی جدید

[1] Mel Frequency Cepstral Coefficients
[2] Classifier

چیست. با این مقدمه مشخص شد که در بحث بازشناسی گفتار نیز ما با این مسئله روبرو هستیم که گفتار ورودی باید با داده های آموزش و از طریق یک طبقه بندی کننده مقایسه شود تا مشخص شود گوینده چه مطلبی را ادا کرده است.

در بازشناسی گفتار سعی بر آن است که سیستم را با داده های تمیز یا به عبارت دیگر سیگنال های گفتاری که در محیط با نویز بسیار کم ادا شده اند آموزش دهند. طبیعی است همانگونه که بیان شد آموزش از طریق بردارهای ویژگی صورت می‌پذیرد. در ادامه و به هنگام ورود یک داده جدید این سیستم ها نیاز به یک طبقه بندی کننده دارند که تشخیص دهند آیا این گفتار ورودی شبیه آن چه با آن آموزش دیده اند می‌باشند یا خیر. طبقه بندی کننده های زیادی در بحث شناسایی آماری الگو مطرح می-شوند که از جمله آنها می‌توان به شبکه های عصبی مصنوعی (ANN[1]) یا مدل مخلوط گوسی (GMM[2]) اشاره نمود. اما در بحث گفتار طبقه بندی کننده بسیار مطرح و کارآیی که مورد استفاده قرار می‌گیرد مدلهای مخفی مارکف یا HMM[3] است. دلیل استفاده از این طبقه بندی کننده ساختار خاص آن است که به یک ویژگی بسیار مهم گفتار منطبق است. این ویژگی پویا بودن سیگنال گفتار در طول زمان می‌باشد. به بیان ساده تر مشخصه مهم گفتار این است ادای یک کلمه حتی از یک گوینده در دو بار پی در پی مشخصه زمانی و فرکانسی مختلفی دارد. از این رو برای تشخیص یک کلمه آنچه اهمیت دارد تشخیص این مطلب است که کلمه از کجا تا کجا ادا شده و احتمال اینکه این سیگنال زمانی کلمه خاصی باشد چقدر است. در واقع این پویا بودن در حوزه زمان است که ما را به سمت استفاده از مدل مخفی مارکف برای بازشناسی می‌برد. مدل مخفی مارکف می‌تواند احتمال یک مسیر خاص را در یک داده گفتاری و فارغ از طول داده محاسبه نماید و با مقایسه این احتمال با آنچه آموزش دیده است تشخیص دهد که چه کلمه یا حتی جمله ای ادا شده است. لازم به ذکر است از HMM می‌توان به منظور استخراج ویژگی نیز استفاده نمود.

[1] Artificial Neural Networks
[2] Gaussian Mixture Models
[3] Hidden Markoves Models

آموزش در بحث بازشناسی گفتار در واقع تنظیم پارامترهای طبقه بندی کننده HMM است. HMM احتمال وقوع بردارها را در حالت ها محاسبه می‌نماید اما آنچه برای ما مهم است این مطلب میباشد که در بین این احتمالات مسیر بهینه کدام است و در انتها احتمال وقوع این مسیر بهینه چقدر است و به چه گفتاری شبیه میباشد. پیدا کردن این مسیر بهینه بر عهده الگوریتم های جستجو است. در بحث گفتار الگوریتم جستجویی که بسیار مورد استفاده قرار میگیرد الگوریتم ویتربی[1] می‌باشد.

مسئله مهمی که گریبان سیستم های بازشناسی گفتار را گرفته است این است که مدل های HMM معمولاً با دادگانی آموزش می‌بینند که در شرایط خاصی ضبط شده اند. مثلاً در شرایط با نویز بسیار کم. بنابر این هنگامی که داده هایی که در شرایط دیگری ضبط شده اند را به سیستم میدهیم، HMM با مشکل مواجه می‌گردد. زیرا جنس سیگنال گفتاری که به آن نشان می‌دهیم متفاوت از سیگنالی است که در حافظه دارد. این مسئله در حضور نویز محیط به شدت حاد می‌شود و دانشمندان مدتهاست به فکر چاره ای برای این مشکل هستند. بهبود دقت بازشناسی گفتار در محیطهای نویزی را مقاوم سازی بازشناسی گفتار می‌گویند. زمینه کاری این رساله نیز رفع این مشکل است.

روش‌های گوناگونی برای مقاوم سازی بازشناسی گفتار ارائه شده اند که می‌توان آنها را به چهار دسته کلی تقسیم بندی کرد. دسته اول روش‌های بهسازی سیگنال گفتار است. هدف از این روش‌ها کم کردن نویز سیگنال به منظور کاهش عدم تطبیق بین داده های آموزش و آزمایش می‌باشد. روش‌های جبران ویژگی یا استخراج ویژگی های مقاوم، روش‌های تطبیق مدل و روشهای مبتنی بر ساختار گوش انسان از دیگر راهکارهای ارائه شده توسط محققین است. ما نیز توجه خود را معطوف به ارائه یک روش پیشنهادی نموده ایم که با بررسی زیرباندهای فرکانسی بردار ویژگی مقاوم تری را نسبت به بردارهای موجود از سیگنال گفتار استخراج نماید.

در ادامه و در فصل دوم مروری بر روش‌های مطرح مقاوم سازی بازشناسی گفتار خواهیم داشت. فصل سوم به توضیحات در باب دادگان و نرم افزارهای مورد استفاده در این پروژه می‌پردازد. روش‌ها و

[1] Viterbi

ایده های پیشنهادی خود را در فصل چهارم تشریح نموده ایم و در ادامه نتایج آزمایش‌ها را در فصل پنجم مشاهده خواهید کرد. در انتها جمع بندی،نتیجه گیری و پیشنهادات برای ادامه کار در فصل ششم آورده شده است.

فصل دوم

مروری بر روش‌های مقاوم سازی بازشناسی گفتار

فصل دوم

مروری بر روش‌های مقاوم سازی بازشتاسی گفتار

یکی از مشکلات اصلی در سیستم‌های بازشناسی گفتار افت کارآیی آن در محیط‌های نویزی است. از آنجاییکه معمولاً آموزش سیستم‌های بازشناس گفتار در محیط‌های به اصطلاح تمیز یا با نویز بسیار کم انجام می‌شود لذا این سیستم‌ها هنگام کار در محیط‌های نویزی با مشکل روبرو می‌شوند. نویزهای محیطی را معمولاً به دو دسته نویز کانال و نویز جمع پذیر تقسیم می‌کنند. از طرفی نویز می‌تواند ایستان[1] یا نا ایستان باشد. تلاش محققین بر آن است روش‌هایی ارائه دهند که جامعیت مطلوبی را دارا باشد به این معنا که تاثیر نویز را جدا از ایستان یا نا ایستان بودن و جمع پذیر یا کانال بودن کاهش دهند.

در طی سالیان اخیرایده ها و روش های متعددی برای مقاوم سازی بازشناسی گفتار ارائه شده اند.این رویکردها را می توان به چهار دسته کلی تقسیم کرد.گروه اول مربوط به روش های کاهش نویز یا به عبارتی بهسازی[2] گفتار است. هدف در این گونه روش‌ها کم کردن نویز سیگنال گفتار به منظور کاهش عدم تطبیق شرایط آموزش و آزمایش قبل از استخراج ویژگی است. در دسته دوم سعی برآن است که

[1] Stationary
[2] Speech Enhancement