سیزهمین کنفرانس مهندسی برق ایران

۲۰-۲۲ اردیبهشت ۱۳۸٤

ترکیب روشهای مبتنی بر مدل و پردازش چندباندی گفتار برای مقاوم سازی بازشناسی گفتار نسبت به نویز

بابک ناصرشریف – دانشگاه علم و صنعت ایران Nasser_s@iust.ac.it

محمدمهدی همایونپور - دانشگاه صنعتی امیرکبیر Homayon@ce,aut.ac.ir

احمد اکبری – دانشگاه علم و صنعت ایران Akbari@just.ac.ir

چکیده: سیستمهای بازشناسی چندباندی گفتار که بر اساس مکانیزم شنوایی انسان عمل می کنند، نرخ بازشناسی را نسبت به سیستم تمام باند به ویژه در حضور نویز بهبود می بخشند. در بازشناسی چندباندی گفتار، سیگنال گفتار ابتدا به چند زیرباند فرکامسی تقسیم می شود و پس از استخراج بردارهای ویژگی از هر زیرباند، این بردارها یا احتمال تخمینی برای آنها با یکدیگر ترکیب می شوند. در کار حاضر سیستم چندباندی بازشناسی گفتار بر مبنای ترکیب ویژگیها مد نظر قرار گرفته است و ترکیب این شیوه با یک شیوه مبتنی بر مدل موسوم به معیار تصویردهی وزن دار پیشنهاد گردیده است. نتایج آزمایشها نشان می دهند که علاوه بر بهتر بودن کارآیی شیوه ترکیب ویژگیها نسبت به

سیستم تمام باند، روش پیشنهادی نیز سبب بهبود چشمگیر کارآیی روش ترکیب ویژگیها می گردد.

کلمات کلیدی: باشناسی چندباندی گفتار، زیرباند، ترکیب ویژگیها، تبدیل موجک، معیار تصویردهی وزن دار

۱-مقدمه

مسئله مقاوم سازی سیستمهای بازشناسی گفتار در برابر نویز را می توان به صورت کاهش میزان عدم تطبیق میان شرایط آموزش و آزمون سیستم درنظر گرفت. روشهایی را که برای کاهش این عدم تطبیق بکار یم روند، می توان به سه گروه اصلی تقسیم کرد: روشهای مبتنی بر داده، روشهای مبتنی بر مدل و شیوه های پردازش چندباندی. روشهای مبتنی بر داده تلاش می کنند تا تاثیرات نویز را بر سیگنالهای گفتار یا ویژگیهای آن کاهش دهند، حال آنکه روشهای مبتنی بر مدل بحای خود سیگنال گفتار یا ویژگیهای آن مدل آلکوستیک گفتار را اصلاح می نمایند. شیوه پردازش چندباندی معمولاً در مورد نویزهایی بکار گرفته می شود که سبب تخریب بخشی از طیف فرکانسی سیگنال گفتار می شوند. در شیوه بازشناسی چندباندی، گفتار تمام باند به چندین زیرباند فرکانسی تقسیم می شود و پس از استخراج بردارهای ویژگی از هر زیرباند، بردارهای ویژگی زیرباندها یا احتمال تخمینی برای آنها توسط بازشناس متناظر

با هر زیرباند، با یکدیگر ترکیب می شوند و به این ترتیب پاسخ بازشناسی بدست می آید.روشهای مبتین بر داده را می توان معمولاً به دو گروه عمده تقسیم کرد: شیوه بهبود گفتار و روشهای جبران ویژگی. شیوه های بهبود گفتار مستقیماً با سیگنال نویزی گفتار سر و کار دارند و با تخمین سیگنال تمیز از سیگنال نویزی در جهت کاهش میزان عدم تطبیق تلاش می کنند. روش تفاضل طیف و آستانه گذاری ضرایب تبدیل موجک سیگنال گفتار نمونه هایی از این دسته هستند. روشهای جبران ویژگی معمولاً عدم تطبیق را به دو طریق کاهش می دهند. در طریق اول، یک تبدیل به ویژگیها اعمال یم شود تا اثر نویز از آنها حذف گردد. تفاضل میانگین ضرایب کپسترال (CMS) و RASTA PLP از جمله چنین روشهایی هتسند. در طریق دیگر، ویژگیهای جدیدی استخراج می شوند که نسبت به تاثیرات نویز مقاوم باشند، همانند ویژگیهای خود همبستگی فاز.

روشهای مبتنی بر مدل، مدل آماری محیط را به نحوی اصلاح می کنند که با شرایط جدید محیطی همانند شرایط نویزی تطبیق یابد. در این تطبیق هیچ نوع فرض یا دانش خاصی در باره خود سیگنال گفتار لازم نیست. این روشها معمولاً نیازمند آموزش برون خط برروی دادگان گفتار نویزی هستند. به عنوان

نمونه ای از این روشها می توان به ترکیب موازی مدلها (PMC) و بازگشت خطی با بیشترین شباهت (MLLR) اشاره کرد.

در بازشناسی چند باندی گفتار، ابتدا سیگنال به چند باند فرکانسی تقسیم می شود. به این ترتیب می توان بخشهای تخریب شده طیق گفتار را از دیگر بخشهای طیف جدا کرد. سپس یک بردار ویژگی از هر زیرباند استخراج می شود که زیربردار ویژگی نامیده می شود. دو روش برای برخورد با این زیربردارها وجود دارد. در روش اول می توان آنها را در کنار یکدیگر قرار داد و به عنوان جایگزینی برای ویژگیهای اصلی استفاده نمود که این شیوه ترکیب ویژگیها نامیده می شود. در روش دیگر زیربردارهای ویژگی بوسیله بازشناس مجزای متناظر یا هر زیرباند، مورد پردازش قرار می گیرند و احتمالی برای آنها تخمین شده می شود و این احتمالات به شیوه خطی یا غیرخطی با یکدیگر ترکیب می شوند. این شیوه ترکیب احتمالات یا ترکیب مدلها نامیده می شود.

در کار حاضر، ما ترکیبی از روشهای مبتنی بر مدل و بازسازی چندباندی گفتار را برای بهبود کارآیی روش بازشناسی مقاوم چندباندی گفتار ارائه می کنیم. در این مقاله، سیستم ترکیب ویژگیها در بازشناسی چندباندی گفتار مد نظر قرار گرفته است و با بکاربردن یک روش مبتنی بر مدل موسوم به معیار تصویردهی وزن دار (WPM)، کارآیی آن بهبود داده شده است. ساختار ادامه مقاله به این

صورت است. بخش دوم به بررسی اصول بازشناسی چندباندی گفتار و ترکیب ویژگیها می پردازد. در بخش سوم چگونگی استفاده از تبدیل موجک برای تقسیم سیگنال گفتار به زیرباندهای فرکانسی شرح داده می شود. بخش چهارم معیار تصویردهی وزن دار را بررسی می کند. در بخش پنجم نتایج آزمایشها ذکر می شود. بخش ششم نیز به جمع بندی و نتیجه گیری کلی اختصاص دارد.

۲-بتزشماسی چندباندی گفتار

چنانکه گفته شد، روشهای بازشناسی چندباندی گفتار به دو دسته کلی تقسیم می شوند: ترکیب ویژگیها و ترکیب احتمالات. ترکیب ویژگی زیرباندها از طریق قراردادن زیربردارهای ویژگی در کنار یکدیگر، ابتدا توسط Okawa در سال ۱۹۹۸ پیشنهاد گردید. الحاق زیربردارهای ویژگی به یکدیگر، یک بردار ویژگی را ایجاد می کند که می توان آن را با شیوه های استاندارد پردازش تمام باند مدل کرد. این امر سبب می شود که همبستگی ممکن میان زیربردارهای ویژگی در مدل آکوستیک درنظر گرفته شود که معمولاً مدل را نیرومندتر و قابل اعتمادتر می کند. مزیت دیگر این ترکیب آن است که پردازش جداگانه برروی زیربردارهای ویژگی نظیر ناهمبسته سازی و دیگر تبدیلات، سبب می شود نویز از یک مجموعه ویژگی تخریب شده به دیگری سرایت نکند. اگرچه این شیوه ترکیب ساده است، لیکن امکان وزن دهی به زیرباندها بر اساس قابلیت اعتماد و

میزان اطلاعات آنها را دارا نیست که این امر نقطه ضعف این شیوه محسوب می گردد.

در روش ترکیب احتمالات با هر زیرباند فرکانسی همانند یک منبع مجزای اطلاعاتی رفتار می شود. پس از عملیات پیش پردازش و استخراج زیربردارهای ویژگی از هر زیرباند، خروجیهای احتمالی کلاس بندهای مربوط به هر زیرباند، در سطحی از تقسیم بندی زمانی با یکدیگر ترکیب می شوند، همانند ترکیب در سطح واج یا ترکیب در سطح هجا، کلمه یا جمله، بسته به نوع کلاس بندی های بکار رفته در زیرباندها شیوه های آماری این ترکیب نیز تغییر می کند. ترکیب احتمالات ممکن است به صورت خطی و با استفاده از یک تابع وزن دهی برای تعیین قابلیت اعتماد نسبی و میزان اطلاعات موجود در هر زیرباند صورت گیرد و یا به شیوه غیرخطی بوسیله ابزارهایی همچون شبکه عصبی MLP انجام شود.

ما در کار پیشین خود سیستم ترکیب احتمالات را مورد بررسی قرار دادیم. در کار حاضر سیستم ترکیب ویژگیها مد نظر قرار گرفته است و با ترکیب آن شیوه مبتنی بر مدل، کارآیی این روش بهبود یافته است.

۳-تقسیم گفتار به زیرباندهای فرکانسی با استفاده از تدبیل موجک

ویژگی اصلی تبدیل موجک بهره جستن از پنجره های زمانی با طول متفاوت بریا باندهای فرکانسی مختلف است. به این ترتیب با استفاده از تبدیل موجک می توان به دقت فرکانسی بالا در باندهای فرکانسی پایین و دقت فرکانسی پایین در باندهای فرکانسی بالا دست یافت. بنا به خصوصیت، تبدیل موجک ابزار قدرتمندی برای مدل کردن سیگنالها ناایستانی همانند سیگنال گفتار است که داریا تغییرات آرام در فرکانسهای پایین و تغییرات ناگهانی در فرکانسهای بالا می باشد. به علاوه مدل فیزیکی حلزون گوش نشان می دهد که حلزون گوش همانند یک تبدیل موجک پیوسته عمل می کند و در آن هر یک از بخشهای مختلف غشای پایه به یک محرک فرکانسی متفاوت عمس العمل نشان می دهد. با توجه به این خصوصیات، ما از تبدیل موجک برای تقسیم گفتار به باندهای فرکانسی استفاده می کنیم.

تبدیل موجک یک سیگنال را می توان به صورت یک نمایش درختی از فیلترهای پایین گذر و بالاگذر در نظر گرفت که در آن هر فیلتر با کاهش نرخ نمونه برداری با ضریب دو دنبال می شود. در تبدیل موجک گسسته تنها شاخه های مربوط به فیلترهای پایین گذر گسترش می یابند، درحالیکه در تبدیل موجک بسته ای درخت بطور کامل در هر دو شاخه پایین گذر و بالاگذر گسترش می یابد. برای بیان بهینه سیگنال با استفاده از تبدیل موجک بسته ای، می توان

درخت تبدیل را با استفاده از معیارهایی همچون کمینه شدن آنتروپی به هنگام توسعه درخت، هرس نمود. نمونه ای از درختهای تبدیل موجک گسسته و تبدیل موجک بسته یا در شکل نشان داده شده اند. در شکل DL و DH به ترتیب نشانگر فیلترهای پایین گذر و بالاگذر هستند.

یکی از مسائل مهم در بازشناسی چندباندی گفتار تعیین تعداد زیرباندها و محدوده فرکانسی زیرباندهاست. ما در کار پیشین خود، نشالن داده ایم که انتخاب باندهای فرکانسی با پهنای باند نامساوی کارآیی بهتری از انتخاب باندهای فرکانسی با پهنای باند مساوی داراست. این مطلب با نحوه عملکرد گوش نیز سازگار است. از این رو در کار حاضر از تبدیل موجک گسسته برپا تقسیم سیگنال ورودی گفتار به چهار باند با پهنای باند نامساوی استفاده گردیده است. تعداد زیرباندهای فرکانسی بنا به نتایج کارهای سایر محققین، برابر چهار در نظر گرفته شده است. علاوه بر اینکه برای تثبیت این مطالب در کار حاضر نیز، مقایسه ای بین تعداد مختلف زیرباندهای نیز صورت گرفته است.

٤-معیار تصویر دهی وزن دار

تئوری معیار تصویردهی وزن دار (WPM) بر اساس این مشاهده Mansour و Juang استوار است که اندازه بردارهای ویژگی کپسترال در حضور نویز

جمع پذیر سفید کاهش می یابد. طبق این خصوصیت در یک معیار محاسباتی بر پایه عمل تصویر کردن معرفی شد که کارآیی بازشناسی گفتار بوسیله روش DTW را در حضور نویز به طور قابل توجهی بهبود بخشید. Carlson و Celemnt این معیار تصویردهی را گسترش دادند و آن را در یک سیستم بازشناسی مبتنی بر مدل مخفی مارکف با چگالی پیوسته (CDHMM) بکار گرفتند. آنها یک عامل مقیاس را در توزیع حالات CDHMM و به عبارت دیگر در توزیع گاوسی احتمالات شباهت دخالت دادند تا به این وسیله کاهش اندازه بردار کریستال جبران گردد. ویژگیهای مورد استفاده آنان ضرایب مل کپستروم (MFCC) بود. نتایج کار آنان نشان داد که بکار بردن این معیار تصویردهی وزن دار در یک سیستم وابسته به گوینده، نرخ بازشناسی کلمات مجزا را به طور قابل توجهی در حضور انواع مختلف نویز از جمله نویز سفید و نویز رنگی بهبود بخشیده است. رابطه جبران تعریف شده آنها برای توزیع گاوسی را می توان به صورت زیر بیان کرد:

که پارامترهای موجود در آن اینگونه تعریف یم شوند:

c_1: بردار مشاهده کپسترال برای قالب t

n: بعد بردار مشاهده

μ_μ: بردار میانگین مخلوط گاوسی jام در حالت i

C_μ: ماتریس کوواریانس مخلوط گاوسی jام در حالت i

λ_μ: عامل مقیاس برای قاب t در مخلوط گاوسی jام در حالت i

$b_{j,i}(c_i)$: احتمال تولید شده برای بردار مشاهده c_i توسط مخلوط گاوسی jام در حالت i

با گرفتن لگاریتم از رابطه (۱) و محاسبه لگاریتم احتمال شباهت، می توان به رابطه ای برای الگوریتم ویتربی دست یافت که بیانگر معیار تطبیق بردار مشاهده و بردارهای میانگین مخلوط گاوسی است. این رابطه را می توان اینگونه نوشت:

با گرفتن مشتق از رابطه (۲) نسبت به λ_μ می توان مقدار بهینه ای برای λ_μ بدست آورد که به واسطه آن احتمال مشاهده $b_{ij}(c_i)$ بیشینه شود. این مقدار بهینه به صورت زیر بدست می آید:

با جایگذاری این مقدار برای λ_μ در رابطه (۲)، مقداری برای لگاریتم احتمال شباهت بدست می آید که WPM نامیده می شود. رابطه (۲) و (۳) برای تمامی مخلوطهای گاوسی موجود در یک حالت مدل مخفی مارکف بکار می روند و احتمال شباهت نهایی با استفاده از مخلوطهای گاوسی تطبیق یافته با معیار تصویردهی وزن دار محاسبه می شود.

در کار حاضر، پس از تشکیل بردار ویژگی با الحاق زیبر بردارهای ویژگی، از روش WPM برای تطبیق بردار میانگین توزیع گاوسی حالات مختلف مارکف با بردار ویژگی نویزی استفاده می شود. به این ترتیب یک مقدار λ_μ برای بردار ویژگی مرکب محاسبه می شود که نوع و نحوه تغییرات زیربردارهای ویژگی در اثر نویز را درنظر می گیرد. به این ترتیب یک مرحله تطبیق با نویز در سطح مدل صورت می گیرد که سبب مقاوم سازی بیشتر بازشناسی با استفاده از شیوه ترکیب ویژگیها می گردد.

۵-آزمایشها و نتایج

نتایج در کار حاضر برروی دادگان TIMIT برای بازشناسی کلمات مجزا گزارش شده اند. این دادگان حاوی ٦٣٠٠ جمله انگلیسی است که توسط ٦٣٠ گوینده و با ٨ لهجه معمول آمریکای شمالی بیان شده اند. در مجموع ٢٤٣٢ جمله متمایز در TIMIT موجود است که شامل ٢ جمله مشترک میان تمامی گویندگان، ٤٥٠ جمله مشترک میان گروههای ٧ نفری گویندگان و ١٨٩٠ جمله تک گوینده است. بانک اطلاعاتی TIMIT به دو بخش آموزش و آزمون تقسیم شده است که بهش آموزش شامل ٤٦٢ گوینده و بخش آزمون شامل ١٦٨ گوینده است و گویندگان و جملات ادا شده در هر یک از این دو بخش با یکدیگر متفاوتند.

در کار حاضر، ۲ جمله مشترک ادا شده بین گویندگان در دادگان TIMIT با دو لهجه (از میان ۸ لهجه) انتخاب شده و با استفاده از برچسب زمانی جملات به کلمات تجزیه گردیده اند. به این ترتیب، ۲۱ کلمه بدست آمده است که بوسیله ۱۵۱ گوینده شامل ۱۰۲ گوینده مرد و ٤۹ گوینده زن ادا شده اند. از این میان، ۲٤۳۹ گویش از ۱۱٤ گوینده برای مجموعه داده آموزش درنظر گرفته شده است. مجموعه داده آزمون نیز حاوی ۷۷۷ گویش از ۳۷ گوینده است. نرخ نمونه برداری هر نمونه گفتاری نیز ۱٦ کیلوهرتز است. سیستم بازشناسی، مدل مخفی مارکف با چگالی پیوسته (CDHMM) است که دارای ٦ حالت است و در هر حالت نیز ۸ مخلوط گاوسی موجود است. آموزش مدل نیز با استفاده از گفتار تمیز (موجود در مجموعه داده آموزش) صورت می گیرد. سه نویز جمع پذیر در کار حاضر مورد استفاده قرار گرفته اند: نویز صورتی (Pink)، نویز محیط کارخانه (factoryl) و نویز سفید که از دادگان نویز NOISEX92 انتخاب گردیده اند. این سه نوع نویز به هر دو مجموعه داده آموزش و آزمون اضافه شده اند. گفتار ورودی با استفاده از تبدیل موجک گسسته با تابع پایه Daubechi مرتبه پنجم (db5) در سه حالت به زیرباندها تقسیم می شود: سه زیرباند با پهنای ۲-۰، ٤-۲ و ۸-٤ کیلوهرتز، چهار زیرباند به پهنای ۱-۰، ٤-۲ و ۸-٤ کیلوهرتز، شش زیرباند با پهنای ۱-۰، ۲-۱، ۳-۲، ٤-۳، ٦-٤ و ۸-٦

کیلوهرتز. سپس با توجه به توزیع بانک فیلتر MEL در روی زیرباندها، از هر زیرباند ضرایب مل کپستروم (MFCC) و مشتق آنها استخراج می شود. در کار حاضر تعداد فیلترها و ضرایب در هر زیرباند این گونه است: در مورد ۳ زیرباند، ۸ فیلتر MEL و ٤ ضریب MFCC و ٤ ضریب مشتق مرتبه اول آن، در مورد ٤ زیرباند، ٦ فیلتر و ۳ ضریب MFCC و ۳ ضریب مشتق مرتبه اول آن، در حالت ٦ زیرباند، ٤ فیلتر و ۲ ضریب MFCC و ۲ ضریب مشتق مرتبه اول آن. این ضرایب از قابهای ۳۰ میلی ثانیه با همپوشانی ۲۰ میلی ثانیه استخراج گردیده اند. طول هر قاب در هر زیرباند با توجه به میزان کاهش فرکانس نمونه برداری در آن زیر باند تعیین گردیده است. در سیستم تمام باند نیز ۱۲ ضریب MFCC و ۱۲ ضریب مشتق اول آن (در مجموع ۲٤ ضریب) از هر قاب استخراج شده است.

شکل (۲) نشانگر خطای بازشناسی کلمه در حضور سه نویز جمع پذیر صوتی، محیط کارخانه و سفید برای تست سیگنال به نویز 10 dB است. این نتایج برای ۳۲۱٦ گویش موجود در دادگان نویزی آموزش و آزمایش بدست آمده اند. در شکل واژه Full نشانگر نتیجه بازشناسی تمام باند است. واژه های FC3,FC4,FC6 بیانگر شیوه ترکیب ویژگیها با استفاده از ٦، ٤ و ۳ زیرباند

هستند و FC4+WPM بیانگر ترکیب معیار تصویردهی وزن دار و شیوه ترکیب ویژگیها با استفاده از ٤ زیرباند است.

چنانکه در شکل مشاهده می گردد، شیوه ترکیب ویژگیها در حضور هر سه نوع نویز سبب بهبود نرخ بازشناسی می گردد. و از این میان ترکیب ویژگیها با استفاده از ٤ زیرباند (FC4) نتایج بهتری از ترکیب در ۳(FC3) و ٦ زیرباند (FC6) داراست. میزان بهبودی نسبی در مورد FC4 برای سه نوع نویز صورتی، محیط کارخانه و سفید به ترتیب برابر ٥/٤٪، ٣/٦٪ و ٤/٨٪ است. با بکارگیری WPM برای شیوه FC4 که نتایج بهتری را داراست، کارآیی این روش به طرز چشمگیری بهبود می یابد. میزان این بهبود نسبت به نتایج تمام باند برای سه نوع نویز مذکور به ترتیب برابر ۱۷/٥٪، ۱۱/۸٪ و ۱۱/٥٪ است.

شکل (۳) خطاب بازشناسی کلمه در حضور سه نویز جمع پذیر صورتی، محیط کارخانه و سفید را برای نسبت سیگنال به نویز 0 dB نشان می دهد. در این مورد نیز بکارگیری شیوه ترکیب ویژگیها سبب کاهش نرخ خطای بازشناسی می شود و ترکیب ویژگیها در ٤ زیرباند کاهش خطای بیشتری را نشان می دهد. میزان نسبی این کاهش برای FC4 در نویزهای صورتی، محیط کارخانه و سفید به ترتیب برابر ۹/٥٪، ٦/۱٪ و ۷/۲٪ است.

کاهش نرخ خطای بازشناسی در اثر ترکیب WPM با روش FC4 (نسبت به بازشناسی تمام باند) برای نویزهای مذکور نیز به ترتیب برابر ۲۸/۶٪، ۳۳/۸٪ و ۱۶/۳٪ است.

٦-جمع بندی و نتیجه گیری

در کار حاضر یک سیستم بازشناسی چندباندی گفتار بر مبنای ترکیب ویژگیها مد نظر قرار گرفت. گفتار ورودی با استفاده از تبدیل موجک گسسته به ۳، ٤ و ٦ زیرباند با پهنای باند فرکانسی نامساوی تقسیم گردید و از هر زیرباند یک زیربردار ویژگی شامل ضرایب MFCC و مشتق آنها، استخراج گردید. یک بردار ویژگی مرکب با الحاق زیربردارهای ویژگی به یکدیگر تشکیل شد و با توجه به نتایج بهتر در ٤ زیرباند، شیوه WPM برای تطبیق بردار ویژگی مرکب حاصل از ٤ زیرباند با میانگین مخلوطهای گاوسی مخفی مارکف بکار رفت. به این ترتیب با اضافه شدن یک مرحله دیگر برای تطبیق با نویز، امکان مقاوم سازی بیشتر روش ترکیب ویژگیها نیز فراهم آمد. نتایج حاصل نشان دادند که روش پیشنهادی، سبب بهبود چشمگیر نرخ بازشناسی کلمه در حضور انواع مختلف نویز پهن باند، باند محدود، ایستان و ناایستان به ویژه در نسبتهای سیگنال به نویز پایین می گردد. در کارهای آینده، بهینه سازی و تعمیم شیوه

WPM برای ادغام با شیوه های ترکیب ویژگیها و ترکیب احتمالات در بازشناسی چندباندی گفتار مد نظر قرار خواهد گرفت.