بازشناسی ارقام فارسی صفر تا نه با استفاده از تصاویر آکوستیک بر پایه ضرایب مل-کپستروم و شبکه عصبی

مسلم خانه بابائی[1]، علی سلیمانی ایوری[2]

[1] گروه الکترونیک، دانشکده مهندسی برق و رباتیک، دانشگاه صنعتی شاهرود، شاهرود، ایران، khanebabaei@shahroodut.ac.ir

[2] دانشیار و عضو هیئت علمی دانشکده مهندسی برق و رباتیک، دانشگاه صنعتی شاهرود، شاهرود، ایران، solimani_ali@shahroodut.ac.ir

چکیده

در این مقاله ابتدا پایگاه داده اعداد صفر تا نه فارسی با استفاده از صدای ۵۰ نفر زن و مرد در محیط ضبط و جمع‌آوری گردیده است. در روش پیشنهادی ابتدا سیگنال پیش‌پردازش شده را قاب‌بندی می‌کنیم و سپس از پنجره بهبودیافته عبور می‌دهیم، در گام بعدی وارد بلوک تبدیل فوریه می‌گردد. حال طیف تبدیل فوریه به بانک فیلتر گوسی داده می‌شود و بعد از آن طیف توان خروجی فیلتر بانک گوسی از تابع ریشه (Root Function) عبور داده شده و سپس با اعمال تبدیل کسینوسی جهت فشرده کردن مولفه‌ها، ضرایب مل-کپستروم به دست می‌آید. در مرحله آخر، تصویر آکوستیک به عنوان ماتریس حاوی ویژگی‌های زمانی و فرکانسی سیگنال گفتار با استفاده از تبدیل معکوس فوریه دوبعدی از ماتریس ضرایب مل-کپستروم تشکیل داده می‌شود. برای طبقه‌بندی و آزمایش داده‌ها، ویژگی‌های به دست آمده با استفاده از یک الگوریتم بهبودیافته در شبکه عصبی پرسپترون با دو لایه پنهان، آموزش داده می‌شوند و در قسمت پایانی میزان نرخ بازشناسی گزارش می‌شود. نتایج آزمایش برای سیگنال به نویزهای متفاوت، نشان دهنده‌ی بهبود نرخ تشخیص سیگنال نویزی توسط روش پیشنهادی است، بطوری که نرخ بازشناسی الگوریتم ارائه شده در حالت بدون نویز ۹۸/۸۵ می‌باشد.

واژه‌های کلیدی:

بازشناسی ارقام، تصویر آکوستیک، شبکه عصبی پرسپترون، ضرایب مل-کپستروم، فیلتر بانک گوسی.

۱- مقدمه

یکی از زیرشاخه‌های پردازش سیگنال، پردازش گفتار است. پردازش گفتار شامل سه شاخه اصلی تبدیل متن به گفتار، بازشناسی گفتار و بهسازی گفتار است. هر سیستم تشخیص گفتار نیازمند استخراج ویژگی است که به وسیله آن بتواند گفتار ورودی را به طور صحیح تشخیص دهد. مقاوم‌سازی این ویژگی‌ها و به طور خاص ویژگی ضرایب فرکانسی مل-کپستروم به عنوان رایج‌ترین آن‌ها در برابر نویز، از اهمیت زیادی در بازشناسی گفتار برخوردار است.

چنانچه محدوده‌ی کلمات مورد بازشناسی را به ارقام محدود کنیم، بازشناسی را بازشناسی اعداد می‌نامیم. در اینجا منظور از عدد، مجموعه-هایی از یک یا چند رقم است که بصورت دلخواه بیان شوند. بازشناسی اعداد امروزه به دلیل کاربردهای متعدد آن، از اهمیت بالایی برخوردار می‌باشد. از موارد کاربرد بازشناسی اعداد می‌توان به شناسایی خودکار شماره شناسایی شخصی، کد ملی، شناسایی حساب بانکی یا شماره عضویت کاربران یک سیستم خدمات رسانی، ارتباط با بانک اطلاعاتی از راه دور، شماره‌گیری خودکار تلفن، دادن رمز ورود به کامپیوتر، ثبت نام دانشجویان از طریق تلفن یا اینترنت پس از اعلام شماره دانشجویی یا کد دروس مورد نظر و مانند آن اشاره نمود. تحقیقات فراوانی بر روی بازشناسی اعداد فارسی انجام گرفته است که به تعدادی از آن‌ها اشاره می‌کنیم.

در سیستمی که در ۱۳۷۷پیاده‌سازی گردید، مدل کلمات صفر تا نه ساخته شد و هر کلمه با شش حالت مدل گردید. در آموزش این سیستم، ۲۰۰ نمونه از هر کلمه که توسط تعداد برابر گوینده‌ی زن و مرد بیان شده بود، مورد استفاده قرار گرفت. ضرایب مورد استفاده در این سیستم، ضرایب کپسترال آنالیز پیش گویی خطی بوده است [۱].

تحقیق دیگری که برای بازشناسی ارقام گسسته از طریق تلفن انجام شد، شبکه‌ی پرسپترون چند لایه مورد استفاده قرار گرفت که در آن بردار ویژگی بعدی توسط شبکه‌ی عصبی تخمین زده می‌شد. آموزش این سیستم با استفاده از ترکیبی از الگوریتم برنامه‌ریزی پویا و الگوریتم آموزش شبکه انجام می‌گرفت. این سیستم بر روی پایگاه داده‌ی تلفنی متشکل از اعداد صفر تا نه فارسی آزمایش گردید. ضرایب مورد استفاده در این سیستم، ضرایب کپسترال در مقیاس مل بوده است. نماد لاتین[1] MFCC

[1] Mel Frequency Cepstrum Coefficients (MFCC)

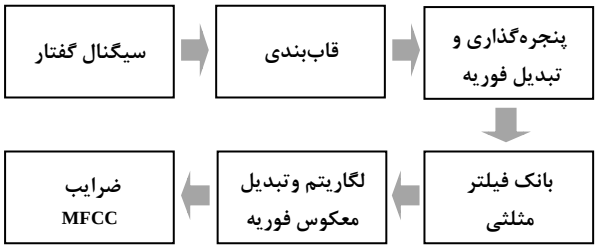
نشان دهنده‌ی این نوع ضرایب است. نتایج بازشناسی در این تحقیق برای داده‌های آزمایشی ۸۱ درصد بود [۲].

در مقاله [۳] که اقدام به بازشناسی گفتار پیوسته فارسی توسط سیستم ترکیبی از مدل مارکوف و شبکه عصبی شد، به راندمان ۷۵ درصد بازشناسی کلمه دست یافتند. در سال ۱۳۹۱ نیز الگوریتمی با استفاده از روش MFCC پیشنهاد شد که عملکرد خوبی در مقابل نویزهای موجود در محیط دارد. در این روش از کپستروم حقیقی (لگاریتم‌گیری)، بانک فیلتر گوسی و یک پنجره بهبودیافته برای بهبود نتایج استفاده شده است [۴].

در این مقاله ابتدا پایگاه داده جمع‌آوری شده معرفی می‌گردد و سپس روش پایه برای استخراج ضرایب مل-کپستروم بیان می‌شود و در مرحله بعد کپسترال در مقیاس مل برای استخراج ویژگی از سیگنال صحبت با استفاده از روش پیشنهادی مورد بررسی قرار می‌گیرد. در بخش سه به معرفی شبکه عصبی و نحوه بازشناسی می‌پردازیم. در قسمت پایانی، نرخ بازشناسی گزارش داده می‌شود و با روش‌های استخراج ویژگی همچون ضرایب مل-کپستروم در روش پایه، ضرایب مل-کپستروم گوسی و ضرایب مل-کپستروم ریشه گوسی برای بازشناسی گفتار در سیگنال به نویزهای مختلف مقایسه گردد. در این مقاله به جای عبارت «ضرایب مل-کپستروم» ،به طور مختصر از « ضرایب مل» استفاده خواهد شد.

پایگاه داده جمع‌آوری شده دارای صدای اعداد صفر تا نه فارسی است که از ۵۰ نفر شامل زن و مرد در بازه سنی ۲۰ تا ۳۰ سال، توسط نگارنده-های این مقاله در محیط ضبط و تهیه گردیده است و جمعاً شامل ۵۰۰ صدای عدد است که با نرم‌افزار Move Maker، به طور یکسان برش داده شده و صداهای هر عدد در فایلی مجزا ذخیره شده است.

روش پایه جهت استخراج ضرایب مل در شکل ۱ نشان داده شده است، در این روش ابتدا قاب سیگنال گفتار از پنجره همینگ عبور داده می‌شود و سپس از بلوک تبدیل گسسته فوریه گذشته و حاصل آن به بانک فیلتر مثلثی اعمال می‌شود. در مرحله بعد خروجی بانک فیلتر مثلثی از بلوک لگاریتم و تبدیل معکوس فوریه گذشته و ضرایب مل را تشکیل می‌دهند.

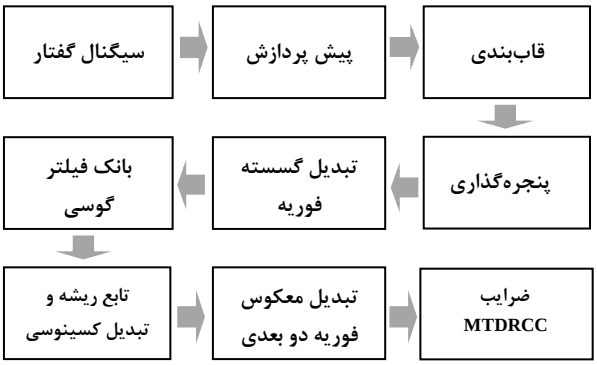


شکل ۱: مراحل استخراج ضرایب مل-کپستروم در روش پایه

در سال ۲۰۰۹ میلادی به جای فیلتر بانک مثلثی از توابع گوسی استفاده شد که به علت ایجاد همبستگی بیشتر بین فریم‌ها موجب بهبود در نرخ تشخیص گشت. نماد لاتین GMFCC[۲] نشان دهنده این نوع ضرایب است [۵]. به طور کلی در سه زمینه مختلف می‌توان جهت بهبود این روش استخراج ویژگی عمل کرد، اولین زمینه بهبود در الگوریتم و بلوک‌های پایه است. دومین زمینه با توجه به اهمیت پیاده‌سازی سخت‌افزاری این الگوریتم، بهبود در قسمت سخت‌افزاری است و نهایتاً سومین زمینه، بهبود و یا ایجاد بلوک‌های جدید و تکمیل کننده در الگوریتم پایه است. روش بهبود در این مقاله از نوع اول و سوم است، یعنی بهبود الگوریتم پایه و تکمیل کننده آن است.

۲- روش پیشنهادی جهت بهبود ضرایب مل

نمودار بلوکی روش پیشنهادی جهت بهبود روش پایه ضرایب مل-کپستروم در شکل ۲ نشان داده شده است.



شکل۲: بلوک دیاگرام روش پیشنهادی استخراج ویژگی

از آنجا که در روش پیشنهادی از کپستروم ریشه و تصویر آکوستیک استفاده شده است، ضرایب حاصل از این روش را با نماد مختصر MTDRCC[۳] نمایش می‌دهیم. در ادامه به شرح مراحل روش پیشنهادی پرداخته می‌شود.

۲-۱- پیش پردازش

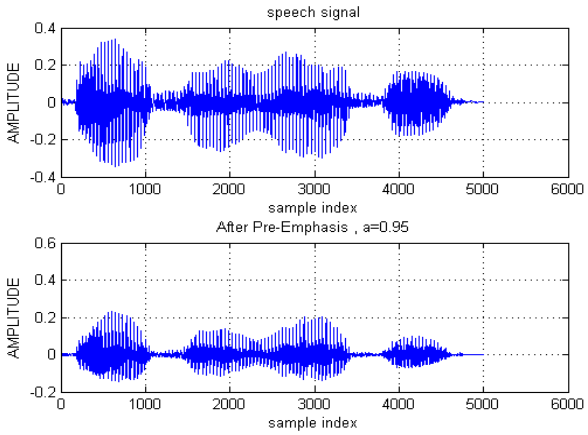
ابتدا سکوت ابتدا و انتهای فایل صحبت را حذف کرده و سپس مقدار DC سیگنال را با استفاده از اختلاف سیگنال گفتار و مقدار میانگین آن جهت بهبود عملیات بعدی نیز حذف می‌نماییم. در مرحله بعد سیگنال گفتار به فیلتر بالاگذر پیش تاکید فرستاده می‌شود. یکی از علل استفاده از فیلتر پیش تاکید این است که این فیلتر اثرات نامناسب حنجره و لب‌ها و تغییرات ناگهانی موجود در سیگنال ناشی از نویزهای محیط را به طور موثری حذف می‌کند و باعث یکنواخت شدن سیگنال گفتار می‌گردد. اگر S(n) سیگنال گفتار و P(n) خروجی فیلتر پیش تاکید باشد، رابطه‌ی (۱) تابع تبدیل این فیلتر را تعریف می‌کند [۱۶].

$$P(n) = S(n) - a.S(n-1) \qquad (۱)$$

پارامتر a ضریب پیش تاکید است و محدوده آن بین ۰/۹ تا ۱ می‌باشد. در شکل ۳ سیگنال گفتار قبل و بعد از اعمال فیلتر پیش تاکید با ضریب پیش تاکید ۰/۹۵ نشان داده شده است.

[2] Gaussian Mel Frequency Cepstrum Coefficients (GMFCC)

[3] Mel-scale Tow Dimension Root Cepstrum Coefficients (MTDRCC)



شکل ۳: سیگنال گفتار قبل و بعد از اعمال فیلتر پیش تاکید

۲-۲- قاب‌بندی

پس از انجام عملیات پیش پردازش، نوبت به مرحله قاب‌بندی سیگنال گفتار می‌رسد. سیگنال گفتار غیر ایستا است، ولی چون اندام‌های گفتار نمی‌توانند سریع‌تر از حد خاصی تغییر یابند، پس در بازه‌های زمانی کوتاه می‌توان آن را ایستا فرض کرد. بنابراین سیگنال گفتار را معمولاً به قاب‌های ۲۰ یا ۲۵ میلی ثانیه‌ای با هم‌پوشانی ۱/۲یا ۱/۳ طول قاب‌ها، تقسیم می‌کنند [۱۵]. هر چه طول قاب‌ها کمتر باشد، با تعداد قاب‌های بیشتری باید سیگنال گفتار را پوشش داد و در نتیجه طول بردار ویژگی استخراج شده بیشتر می‌شود و حجم محاسبات بالا می‌رود، ولی اطلاعات بیشتری از گفتار در اختیار ما قرار می‌گیرد. در اینجا هر قاب را با طول ۵۰۰ نمونه و هم پوشانی ۱۵۰ نمونه در نظر گرفته‌ایم.

۲-۳- پنجره‌گذاری

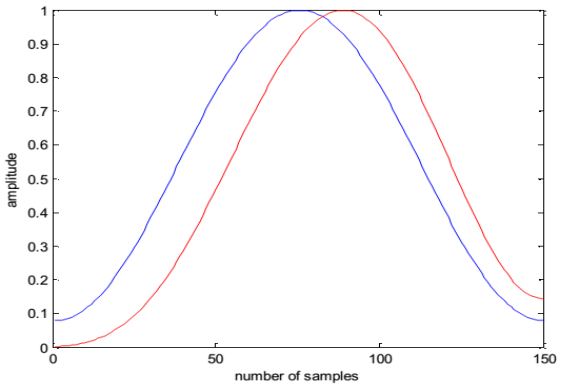
در این مرحله هر قاب به طور جداگانه در یک پنجره ضرب می‌شود تا اثر ناپیوستگی سیگنال در ابتدا و انتهای سیگنال کم شود. انتخاب پنجره بسیار مهم است، زیرا حاشیه‌های یک قاب در کم یا زیاد شدن سیگنال خطا تاثیر دارند، در الگوریتم پایه از پنجره گذاری همینگ W(n) در رابطه (۲) استفاده می‌شود [۱۶]. اما پنجره‌ای که ما در اینجا استفاده می‌کنیم، مطابق رابطه (۳)، پنجره بهبود یافته $W_{New}(n)$ می‌باشد. اگر قاب با X(n) و قاب پنجره‌گذاری شده با $\overline{X}(n)$ نمایش داده شود، اعمال پنجره همانند رابطه (۴) خواهد بود. N تعداد نمونه‌ها در یک قاب و K شماره قاب است.

$$W(n) = 0.54 - 0.46\cos\frac{2n\pi}{N-1}, 0 \le n \le N-1 \qquad (۲)$$

$$W_{New}(n) = nW(n), 0 \le n \le N-1 \qquad (۳)$$

$$\overline{X}_K(n) = X_K(n)W_{New}(n), 0 \le n \le N-1 \qquad (۴)$$

در پنجره بهبود یافته سه شاخص مهم پراکندگی، همگرایی بخش‌های جانبی و عرض بخش اصلی پنجره در نظر گرفته شده است. در این پنجره، نسبت به یک پنجره‌ی همینگ ساده، فاکتور پراکندگی طیفی و نیز عرض وجه اصلی افزایش و فاکتور همگرایی وجه‌های جانبی کاهش می‌یابد که دو مورد اول تغییراتی مطلوب و مورد آخر اثری نامطلوب می‌باشد [۶]. با توجه به بهبود نتایج، می‌توان از اثر نامطلوب چشم پوشی کرد. شکل ۴ تفاوت این دو پنجره را نشان می‌دهد.



شکل ۴: پنجره همینگ (آبی رنگ) و پنجره بهبودیافته (قرمز رنگ)

۲-۴- تبدیل فوریه گسسته

چون نوع صدا به توزیع انرژی سیگنال در حوزه فرکانس مربوط است، به اطلاعات آن در حوزه فرکانس احتیاج داریم. با استفاده از رابطه (۵)، سیگنال x(n) درون هر قاب را به حوزه فرکانس می‌بریم. از آنجا که بیشتر اطلاعات صوت در طیف تبدیل فوریه آن قرار دارد، طیف تبدیل فوریه هر قاب را به دست می‌آوریم. پارامتر Ms، تعداد نقاط در تبدیل فوریه است.

$$X(k) = \sum_{n=1}^{Ms} x(n)\, e^{-j\frac{2\pi nk}{Ms}}, 1 \le k \le Ms \qquad (۵)$$

۲-۵- بانک فیلترگوسی

ضرایب مورد نظر، از اعمال مجموعه‌ای از بانک فیلتر که تمام طیف فرکانسی را پوشش می‌دهند، به دست می‌آیند. در ساختار بانک فیلتر میان‌گذر که مجموعاً کل پهنای باند سیگنال را پوشش می دهند، طیف تبدیل فوریه قاب‌ها از بین فیلترها عبور می‌کند. در بازشناسی گفتار از انواع مختلفی از بانک فیلترها استفاده می‌شود. این فیلترها تفکیک فرکانسی سیستم ادراک گوش انسان را شبیه‌سازی می‌کنند. در الگوریتم پایه محاسبه ضرایب MFCC معمولاً از بانک فیلتر مثلثی استفاده می‌گردد. در این نوع فیلتر بانک، اگر میزان هم‌پوشانی قاب‌ها کافی نباشد، اطلاعات بخش‌هایی از فریم که در نقاط ابتدایی و انتهایی و خارج از زیربخش‌ها قرار می‌گیرند، از دست می‌روند، چون مثلث‌ها در خارج از زیرباندها وزنی ندارند. اما اگر به جای این فیلتر بانک از یک فیلتر بانک گوسی استفاده کنیم، به دلیل وجود وزن در خارج از زیرباندهای آن، مانع از دست رفتن اطلاعات در این بخش‌ها می‌گردد، همچنین در بانک فیلتر گوسی، همبستگی بیشتری بین فیلترهای مجاور وجود دارد و می‌توان این همبستگی را با استفاده از پارامتر آلفا در رابطه (۸) افزایش یا کاهش داد. برای ایجاد بانک فیلترها، ابتدا باید فرکانس‌ها را به حوزه مل که واحد شنیداری گوش انسان هست منتقل کنیم. درحقیقت رابطه (۶) نگاشتی از فرکانس واقعی f به فرکانس مل f_{mel} می‌باشد.

$$f_{mel} = 2595\,\mathrm{Log}_{10}(1+\frac{f}{700}) \qquad (۶)$$

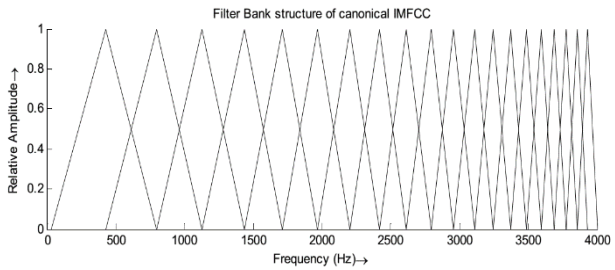
در ادامه پارامتر k_{bi} که در بانک فیلتر مثلثی نقاط مرزی را مشخص می‌کردند، در رابطه (۷) به دست می‌آوریم. همچنین k_{bi}ها در تعیین پارامتر پراکندگی σ_i کاربرد دارند. بعد از آن که واریانس هر زیربخش از بانک فیلتر در رابطه (۸) محاسبه شد، نهایتاً با معادله پایانی در رابطه (۹) فیلتر بانک گوسی $\Psi_i(k)$ را به دست می‌آوریم [۵].

$$k_{bi} = (\frac{Ms}{Fs}).f^{-1}_{mel}[f_{mel,min} + \frac{i.(f_{mel,max} - f_{mel,min})}{Q+1}] \quad (۷)$$

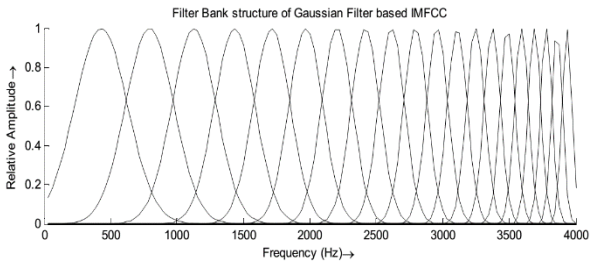
$$\sigma_i = \frac{k_{bi+1} - k_{bi}}{\alpha} \quad (۸)$$

$$\Psi_i(k) = e^{-\frac{(k-k_{b_i})^2}{2\sigma_i^2}} \quad (۹)$$

پارامتر Q تعداد فیلتر بانک و $1 \le i \le Q$ می‌باشد. همچنین Ms تعداد نقاط در تبدیل فوریه گسسته است که منجر به طیف انرژی می‌شود. Fs نیز فرکانس نمونه‌برداری است. پارامتر α تنظیم کننده واریانس می‌باشد، به نحوی که هر چقدر پارامتر آلفا بزرگتر باشد، انحراف معیار فیلترهای گوسی کمتر می‌شود و بلعکس. در این مقاله مقدار آلفا برابر با ۲ است، زیرا در این حالت همبستگی بهتری بین زیرباندهای مجاور در بانک فیلتر گوسی ایجاد می‌شود [۵]. حال می‌توان بانک فیلتر گوسی را برای استخراج ضرایب GMFCC طراحی کرد. شکل ۵ و ۶ به ترتیب نمودارهای بانک فیلتر مثلثی و بانک فیلتر گوسی را نمایش می‌دهد.



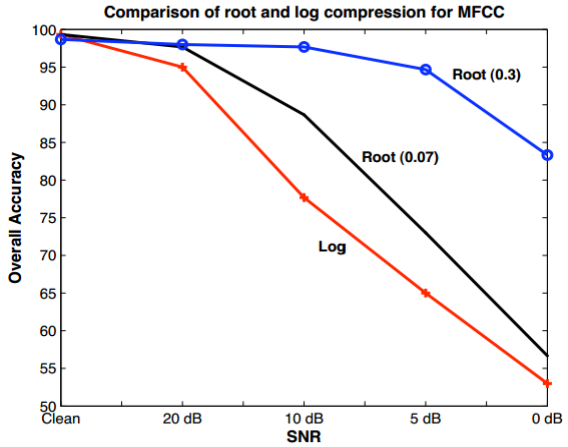
شکل ۵: بانک فیلتر مثلثی



شکل ۶: بانک فیلتر گوسی

۲-۶- تابع ریشه و تبدیل کسینوسی

در الگوریتم پایه MFCC بعد از خروجی بانک فیلتر، لگاریتم اندازه انجام می‌شود که به روش کپستروم حقیقی معروف است. در روش پیشنهادی خروجی بانک فیلتر گوسی را از تابع ریشه عبور می‌دهیم، این روش که کپستروم ریشه نام دارد، باعث خواهد شد که ضرایب مل به دست آمده حرکت نرم‌تری داشته باشند و از این رو در حذف نویز سوار شده بر سیگنال گفتار مفید خواهد بود [۸]. پارامتر گاما در تابع ریشه می‌تواند عددی بین ۱- و ۱ باشد، به صورتی که هر چه گاما به یک نزدیک‌تر باشد، ضرایب مل به دست آمده تغییرات بیشتری خواهند داشت [۷]. در این مقاله برای دستیابی به نتایج بهتر، مقدار پارامتر گاما، ۰/۳ در نظر گرفته شده است. شکل ۷ تغییرات تابع لگاریتم و تابع ریشه را به ازای افزایش نویز یا کاهش سیگنال به نویز نشان می‌دهد.



شکل ۷: مقایسه تابع ریشه و تابع لگاریتم در سیگنال به نویز متفاوت

همانطور که گفته شد، تابع ریشه نسبت به تابع لگاریتم در برابر افزایش نویزهای موجود در محیط مقاوم‌تر است [۱۱]. عملیات کپستروم ریشه با استفاده از رابطه (۱۰) انجام می‌شود. برای کاهش مقدار مولفه‌ها و فشرده‌سازی، بر روی خروجی تابع ریشه، تبدیل کسینوسی گسسته مطابق رابطه (۱۱) اعمال می‌شود.

$$Y_{\gamma,j}(m) = |Y_j(k)|^{\gamma},\ -1 \le \gamma \le 1 \quad (۱۰)$$

$$C_j(n) = \sum_{m=1}^{F} Y_{\gamma,j}(m) \cos(\frac{n(m-0.5)\pi}{F}), 1 \le n \le P \quad (۱۱)$$

$Y_j(k)$ خروجی بانک فیلتر گوسی شماره k برای قاب j، F تعداد فیلترهای گوسی، $Y_{\gamma,j}(m)$ خروجی حاصل از تابع ریشه، $P \le N$ طول بردار ضرایب مل و N طول یک قاب گفتار، گاما پارامتر ریشه کپستروم و C_j بردار ضرایب فرکانسی به دست آمده برای قاب j است. از آنجا که ضرایب مدنظر در رابطه (۱۱) با استفاده از بانک فیلتر گوسی و تابع ریشه حاصل شده اند، این ضرایب را با نماد لاتین[4] GMFRCC نشان می‌دهیم.

[4] Gaussian Mel Frequency Root Cepstrum Coefficients (GMFRCC)

۲-۷- تبدیل معکوس فوریه دو بعدی

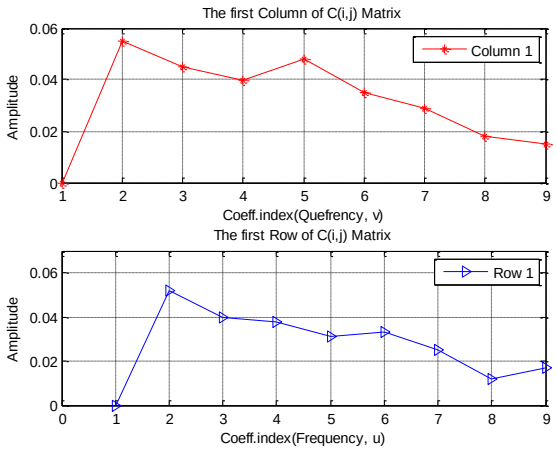
معمولاً ویژگی‌های استخراجی از سیگنال گفتار به صورت تک بعدی و از حوزه زمان یا فرکانس هستند، در صورتی که ویژگی‌های مفید در هر دو حوزه زمان و فرکانس وجود دارند. از این رو ابتدا از کنار هم قرار دادن بردار ضرایب GMFRCC هر قاب، ماتریس Y(i,j) که شامل کل ضرایب GMFRCC سیگنال گفتار هست، تشکیل داده می‌شود. سپس با استفاده از رابطه (۱۲)، به صورت دو بعدی از ماتریس Y تبدیل معکوس فوریه می‌گیریم [۷]. نتیجه این کار ماتریسی است به نام تصویر آکوستیک که دارای دو بعد زمان و فرکانس می‌باشد [۱۲]، به عبارتی ویژگی‌های هر دو حوزه زمان و فرکانس را شامل می‌شود [۸]. از ماتریس تصویر آکوستیک، جهت ایجاد نمودار پرکاربرد اسپکتروگرام نیز استفاده می‌شود.

$$\hat{x}(u,v)=\frac{1}{MP}\sum_{m=1}^{M}\sum_{p=1}^{P}Y(i,j)e^{\frac{j2pv\pi}{P}}e^{\frac{j2mu\pi}{M}} \qquad (۱۲)$$

M تعداد قاب‌های استفاده شده برای محاسبه ماتریس Y و P نشان دهنده طول بردار ضرایب مل در یک قاب است، همچنین $1\le u\le P$ و $1\le v\le M$ می‌باشد. ماتریس ضرایب نهایی بعد از محاسبه قدر مطلق در رابطه (۱۳) به دست می‌آید.

$$C(i,j)=|\hat{x}(u,v)| \qquad (۱۳)$$

در شکل ۸ ضرایب اولین ستون و اولین سطر از ماتریس C(i,j) نمایش داده شده است. طبق شکل ۸، اکثر ویژگی‌های مهم و مفید سیگنال گفتار در ابعاد پایین ماتریس C(i,j) قرار دارد، بنابراین دریه‌های موجود در ابعاد پایین را به عنوان یک زیر ماتریس که حاوی ویژگی‌های مهم سیگنال گفتار است، از ماتریس C(i,j) جدا می‌کنیم [۸]. در این مقاله زیر ماتریس دارای ابعادی برابر با 5×20 می‌باشد، یعنی از هر سیگنال گفتار، ۱۰۰ ویژگی زمانی و فرکانسی استخراج می‌شود.

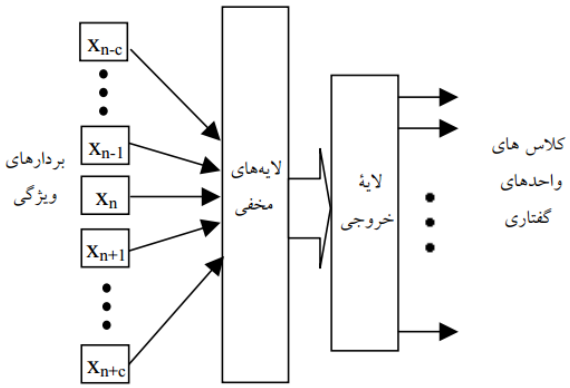


شکل ۸: ضرایب ستون اول و سطر اول ماتریس C(i,j)

۳- طبقه‌بندی با شبکه عصبی مصنوعی

مطالعه شبکه‌های عصبی مصنوعی تا حد زیادی ملهم از سیستم‌های یادگیری طبیعی است که در آن‌ها یک مجموعه پیچیده از نرون‌های به هم متصل در کار یادگیری دخیل هستند. نوعی از یک شبکه عصبی مصنوعی برمبنای یک واحد محاسباتی به نام پرسپترون ساخته می‌شود. یک پرسپترون برداری از ورودی‌های با مقادیر حقیقی را گرفته و یک ترکیب خطی از این ورودی‌ها را محاسبه می‌کند. در این مقاله از یک شبکه پرسپترون با یک لایه ورودی، دو لایه میانی یا مخفی و یک لایه خروجی استفاده شده است. در اینجا ما ۱۰ کلاس داریم. تعداد نرون لایه ورودی به اندازه بردار ویژگی‌ها و تعداد نرون‌های لایه میانی به دلخواه توسط کاربر تعیین شده است. لایه خروجی هم دارای یک نرون می‌باشد. ارتباط بین هر لایه توسط اتصال‌هایی که به نرون‌ها متصل شوند، بر قرار می‌شود. به این اتصال‌ها اصطلاحاً وزن‌های شبکه گویند [۱۳].

لازم به ذکر است علاوه بر وزن‌ها، هر نرون متاثر از کمیت مستقل دیگری با نام بایاس می‌باشد. تاثیر پذیری مقادیر وزن‌ها و بایاس‌ها در نرون‌های هر لایه، به لایه‌ی بعدی توسط تابعی تحت عنوان تابع انتقال مشخص می‌شود. وزن‌های مربوط به هر نرون، عامل مهم و حائز اهمیت در شبکه می‌باشند، لذا سعی بر انتخاب مقادیر مناسب این وزن‌ها که کمترین خطای شبکه را در بر داشته باشد، شده است. به این عمل، آموزش شبکه گفته می‌شود [۹]. شکل ۹ نمایی از ساختار شبکه عصبی استفاده شده را نشان می‌دهد.



شکل ۹: ساختار یک شبکه عصبی پرسپترون با لایه‌های مخفی

در طبقه‌بندی کردن با استفاده از شبکه عصبی مصنوعی برای کاهش خطا در به دست آوردن وزن‌ها، از قانون دلتا استفاده شده است. ایده اصلی این قانون استفاده از گرادیان نزولی برای جستجو در فضای فرضیه وزن‌های ممکن می‌باشد. الگوریتم گرادیان نزولی در فضای وزن‌ها به دنبال برداری می‌گردد که خطا را حداقل کند. این الگوریتم از یک مقدار دلخواه برای بردار وزن شروع کرده و در هر مرحله وزن‌ها را طوری تغییر می‌دهد که خطا کاهش داده شود، این قانون پایه روش انتشار به عقب است که برای آموزش شبکه با چندین لایه مخفی بکار می‌رود [۱۰] [۱۴].

۴- نتایج شبیه‌سازی

به منظور آزمایش کردن روش پیشنهادی و بررسی عملکرد ضرایب MTDRCC در نرخ بازشناسی گفتاری که به نویز آلوده شده است، یک سری آزمایش انجام گرفته است. آزمایش‌های این مقاله بر روی یک پایگاه داده اعداد فارسی بین صفر تا نه شرح داده شده، انجام شده است. درصد

این دادگان برای آموزش شبکه ۷۰ و ۳۰ درصد برای تست و آزمایش شبکه استفاده شده است. ابتدا از هر سیگنال گفتار، چهار ویژگی با نمادهای لاتین MTDRCC، GMFRCC، GMFCC و MFCC استخراج شده است. برای سیگنال گفتار بدون نویز، تاثیر هر کدام از ویژگی‌ها را در نرخ بازشناسی آزمایش می‌کنیم، سپس به تدریج نسبت سیگنال به نویز را کاهش می‌دهیم، برای ایجاد نویز بر روی سیگنال از نویز سفید گوسی استفاده شده است. نتایج حاصل در جدول ۱ نشان داده شده است. با مقایسه خروجی شبکه برای روش پیشنهادی و سه روش دیگر در نسبت سیگنال به نویزهای متفاوت، مشاهده می‌شود که الگوریتم پیشنهادی در مقابل نویز اعمالی مقاوم‌تر است و همچنین درصد تشخیص بالاتری نسبت به روش‌های دیگر دارد.

جدول ۱ : مقایسه نرخ بازشناسی به ازای ویژگی و سیگنال به نویز مختلف

SNR/Feature	MFCC	GMFCC	GMRFCC	MTDRCC
No Noise	85.5	89.55	92.79	98.85
25 dB	83.4	87.56	88.22	97.38
10 dB	78	82.4	85.45	96.26
5 dB	70.45	75.6	76.7	89.19
0 dB	61.52	65.15	69.05	80.09
-5 dB	54.1	64.18	66.4	79.54

۵- نتیجه‌گیری

در این مقاله یک الگوریتم بهبود یافته بر پایه ضرایب مل-کپستروم به منظور افزایش نرخ بازشناسی گفتار آلوده به نویز پیشنهاد شد. استفاده از پنجره بهبودیافته تاثیر قابل توجهی در بهبود نتایج دارد. به دلیل وجود وزن در خارج از زیرباندهای فیلتر بانک گوسی، مانع از دست رفتن اطلاعات در این بخش‌ها می‌گردد و در بهبود الگوریتم تاثیر قابل توجهی دارد، همچنین استفاده از کپستروم ریشه باعث می‌شود که ضرایب مل در مقابل نویزهای اعمالی به سیگنال گفتار مقاوم‌تر باشد. استخراج ویژگی‌های نهایی از ماتریس تصویر آکوستیک، باعث می‌شود که بتوان از ویژگی‌های مهم و مفید موجود در هر دو حوزه زمان و فرکانس سیگنال گفتار استفاده کرد. شبکه عصبی طراحی شده به نحوی است که در آموزش شبکه کمترین خطای موجود و در پی آن مناسب‌ترین وزن‌ها را داشته باشیم. طبق جدول ۱، الگوریتم پیشنهادی در برابر افزایش نویز نسبت به دیگر روش‌ها مصون‌تر است و نرخ بازشناسی بالاتری نسبت به سایر روش‌ها از خود نشان می‌دهد.

۶- سپاسگزاری

در پایان لازم است از تمام کسانی که در تهیه و ضبط پایگاه داده با نگارنده‌های این مقاله همکاری نمودند، کمال تشکر و قدردانی را داشته باشیم.

مراجع

[۱] رستم‌زاده و همکاران، "بازشناسی گفتار فارسی ناپیوسته، به صورت ناوابسته به گوینده به کمک مدل‌های پنهان مارکوف با چگالی پیوسته"، ششمین کنفرانس مهندسی برق ایران، تهران، دانشگاه صنعتی خواجه نصیرالدین طوسی، دانشکده مهندسی برق، ۱۳۷۷.

[۲] همایون‌پور و نجاری، "بازشناسی ارقام فارسی ناوابسته به گوینده با استفاده از مدل پیشگوی عصبی"، هفتمین کنفرانس مهندسی برق ایران، تهران، مرکز تحقیقات مخابرات ایران، ۷۵-۸۱، ۱۳۷۸.

[۳] طاهری و همکاران، "بازشناسی گفتار پیوسته فارسی در دایره لغات متوسط بروش ترکیب شبکه‌های عصبی و مدل‌های مارکوف پنهان"، دهمین کنفرانس مهندسی برق ایران، دانشگاه تبریز، دانشکده مهندسی، ۱۳۸۱.

[۴] مروی و همکاران، "بازشناسی مقاوم گفتار فارسی با استفاده از ضرایب مل-کپستروم بهبودیافته و شبکه عصبی"، یازدهمین کنفرانس سیستم‌های هوشمند ایران، دانشگاه خوارزمی، اسفند ۱۳۹۱.

[5] S. Chakroborty, and G. Saha, "Improved text-independent speaker identification using fused MFCC & IMFCC feature sets based on Gaussian filter," International Journal of Signal Processing, vol. 5, no. 1, pp. 11-19, 2009.

[6] M. Sahidullah, and G. Saha, "A novel windowing technique for efficient computation of MFCC for speaker recognition," IEEE Signal Processing Letters, vol. 20, no. 2, pp. 149-152, 2013.

[7] X. Wang, and Z. Han, "A novel acoustic feature extraction algorithm based on root cepstrum coefficients and CCBC for robust speech recognition," Second International Symposium on Intelligent Information Technology Application, 2008.

[8] H. Marvi, Efficient feature extraction based on two-dimensional cepstrum analysis for speech recognition, University of Surrey, Guildford, UK, December 2004.

[9] A. M. Othman, and M. H. Riadh, "Speech recognition using scaly neural networks," International Journal of Electrical and Computer Engineering, vol. 38, pp. 253-258, 2008.

[10] V. Skorpil, and J. Stastny, "Back-propagation and k-means algorithms comparison," IEEE International Conference on Signal Processing, vol. 3, 2006.

[11] P. Alexander and P. Lock wood, "Root cepstral analysis: A unified view application to speech recognition processing in car noise environments," Speech communication, 1993.

[12] H Marvi, I. Paraskevas, and E. Chilton, " Acoustic classification using time-frequency distributions," In Proceeding of the Institute of Acoustics, volume 26 Pt.2, pp. 612-620, 2004.

[13] J. Esmaily, R. Moradinezhad, J. Ghasemi, " Intrusion Detection System Based on Multi-Layer Perceptron Neural Networks and Decision Tree," 7th International Conference on Information and Knowledge Technology, 2015.

[14] K. Gurney, "An Introduction to Neural Networks," Ucl, Press Limited Taylor & Francis Group London, 1997.

[15] Garima Vyas, Barkha Kumari, "Speaker Recognition System Based On MFCC and DCT," International Journal of Engineering and Advanced Technology (IJEAT), Volume. 2, Issue. 5, June 2013.

[16] Lindasalwa Muda and et al, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques," International Journal of Computing, Volume. 2, Issue. 5, March 2010.