تشخیص خودکار گوینده مبتنی بر ویژگی‌های استخراج شده از بانک فیلتر گابور و شبکه‌های عصبی کانولوشنال

عبدالرضا رشنو[۱]، صادق فدایی[۲،*] و عبدالصمد حمیدی[۳]

اطلاعات مقاله

نوع مقاله: پژوهشی
دریافت مقاله: ۱۴۰۱/۰۱/۰۴
بازنگری مقاله: ۱۴۰۱/۰۵/۰۵
پذیرش مقاله: ۱۴۰۱/۰۶/۲۷

واژگان کلیدی:
بانک فیلتر گابور،
اسپکتروگرام،
تشخیص گوینده،
شبکه‌ی عصبی کانولوشنال.

چکیده

صدای یک انسان حاوی خصوصیاتی از قبیل: قومیت، جنسیت، احساس، سن و اطلاعات دیگری از فرد است و موضوع تشخیص گوینده به شناسایی هویت افراد بر اساس صدای آنها می‌پردازد. اگرچه محققان در طول سال‌های گذشته در این زمینه فعالیت داشته‌اند و روش‌هایی برای بهبود دقت تشخیص گوینده پیشنهاد داده‌اند اما هنوز چالش‌هایی در این زمینه وجود دارد. در این مقاله یک روش جدید تشخیص گوینده مبتنی بر فیلترهای گابور و شبکه‌های عصبی کانولوشنال ارایه شده است. در روش پیشنهادی، ابتدا اسپکتروگرام سیگنال صحبت فرد تشکیل می‌شود. سپس با طراحی موثر فیلترهای گابور، بانک فیلتر گابور ایجاد می‌گردد. در مرحله‌ی بعد اسپکتروگرام سیگنال از بانک فیلتر گابور عبور داده شده و ویژگی‌های سیگنال صحبت استخراج می‌شود. در مرحله‌ی آخر با استفاده از یک شبکه‌ی عصبی کانولوشنال، گوینده شناسایی می‌شود. برای ارزیابی روش پیشنهادی از دو پایگاه داده‌ی Aurora2 و TIMIT استفاده شده است. نتایج نشان می‌دهد که روش پیشنهادی دقت بهتری نسبت به روش‌های پیشین دارد.

۱-مقدمه

انسان‌ها داری ویژگی‌های زیادی هستند که باعث تمایز آنها از یکدیگر می‌شود. بعضی از این ویژگی‌ها می‌تواند به وضوح دیده شود و قابل تشخیص باشد. ویژگی‌های ظاهری و صورت، ویژگی‌های آوایی و رفتارها از این قبیل ویژگی‌ها هستند. برخی دیگر از ویژگی‌ها مانند اثر انگشت، ژن‌ها و یا ساختارهای DNA به راحتی قابل مشاهده نبوده و ظاهری نیستند و باید با اندازه‌گیری‌های پیچیده به هویت آنها پی ببریم. در سال‌های اخیر، بیومتریک به عنوان یک روال علمی و سازمان‌یافته در راستای محاسبه‌ی خودکار ویژگی ها و ساختارهای شخصی یک فرد برای سیستم‌های امنیتی ارائه شده است [۱].

ویژگی‌های صوتی به عنوان یک ابزار بیومتریک به چند دلیل می‌توانند مفید واقع شوند. یکی اینکه صوت هر شخص می‌تواند با ابزار ساده و ارزان ضبط شود و بدون نیاز به حضور او مورد استفاده قرار گیرد و دوم اینکه برای کاربردهای امنیتی و کنترلی از را دور به وسیله تلفن می تواند مورد استفاده قرار گیرد. به طور کلی در یک سیگنال صوتی اطلاعات مختلفی نهفته است. مهمترین این اطلاعات پیامی است که در سیگنال نهفته است و مشخص می‌کند که فرد می‌خواهد با این سیگنال چه پیامی را منتقل کند. اما اطلاعات دیگری که می‌توان از یک سیگنال صوتی بدست آورد شامل هویت گوینده، زبانی که صحبت می‌شود، لهجه گوینده، احساس گوینده و ... است [۲].

* پست الکترونیک نویسنده مسئول: s.fadaei@yu.ac.ir
۱. استادیار، دانشکده مهندسی، دانشگاه لرستان
۲. استادیار، دانشکده مهندسی، دانشگاه یاسوج
۳. استادیار، دانشکده مهندسی، دانشگاه لرستان

https://doi.org/10.22075/jme.2022.26690.2245

تشخیص گوینده عملی است که در آن هویت یک شخص با استفاده از صدای آن تشخیص داده می‌شود و به دلایل امنیتی و کنترل تلفنی از راه دور در سال‌های اخیر مورد توجه قرار گرفته است. تاریخچه‌ی تشخیص گوینده به سال ۱۶۶۰ برمی‌گردد که یک سند نشان می‌دهد هویت شخصی از روی صدایش تعین شده است. این سند قسمتی از دفاعیات یک وکیل مدافع در خصوص مرگ چارلز اول در یکی از جلسات محاکمه است. البته مساله‌ی تشخیص گوینده به عنوان یک زمینه‌ی تحقیق علمی حتی تا دو قرن بعد از آن هم مطرح نشد، ولی بعد از اختراع تلفن و ضبط صوت امکان تشخیص گوینده میسر گردید. بالاخره در سال ۱۹۶۶ یک دادگاه قانونی برای اولین بار تشخیص گوینده را بر اساس طیف گفتار به رسمیت شناخت. مواردی که تشخیص گوینده توسط انسان نباشد و با یک سیستم خاص انجام شود سیستم‌های تشخیص گوینده‌ی خودکار نامیده می‌شود. تحقیق در زمینه‌ی سیستم‌های تشیخص گوینده‌ی خودکار به حدود چهار دهه پیش، یعنی بین سال‌های ۱۹۷۰ تا ۱۹۸۰ برمی‌گرد [۳،۴]. به طور کلی دو نوع سیستم تشخیص گوینده وجود دارد: ۱-وابسته به متن[1] که در آن سیستم تشخیص‌دهنده، دانش اولیه‌ای از متن یا سیگنالی که قرار است گوینده به سیستم بگوید، دارد و به عبارتی گوینده باید همیشه یک متن خاص را به سیستم بگوید. ۲-مستقل از متن[2] که در آن سیستم تشخیص‌دهنده، هیچ دانش اولیه‌ای از متن یا سیگنالی که قرار است گوینده به سیستم بگوید، ندارد و گوینده هر متن دلخواهی را می‌تواند برای سیستم بخواند. سیستم نوع دوم جامع‌تر، طراحی آن سخت‌تر و قابلیت انعطاف آن بیشتر است [۲]. مطالعات نشان می‌دهد فیلتر گابور برای تحلیل ویژگی‌های صوتی مناسب است، لذا در این مقاله، بر اساس ویژگی‌های زمان-فرکانس گابور و شبکه‌های عصبی کانولوشنال(CNN[3]) یک سیستم تشخیص گوینده پیشنهاد شده است. در روش پیشنهادی، ابتدا اسپکتروگرام سیگنال تشکیل می‌شود سپس با استفاده از فیلترهای گابور، ویژگی‌های سیگنال استخراج شده و در نهایت با استفاده از یک شبکه‌ی عصبی کانولوشنال، گوینده شناسایی می‌شود. ادامه‌ی مقاله به این صورت سازمان‌دهی شده است: بخش ۲ به معرفی کارهای مرتبط می‌پردازد و روش پیشنهادی در بخش ۳ ارایه می‌شود. پایگاه‌های داده‌ی مورد استفاده و تنظیم پارامترها در بخش ۴ آمده است. نتایج پیاده‌سازی الگوریتم پیشنهادی در بخش ۵ ارایه می شود و در نهایت مقاله در بخش ۶ جمع‌بندی شده است.

۲-کارهای مرتبط

اگرچه در زمینه‌ی تشخیص گوینده پیشرفت‌هایی صورت گرفته است اما هنوز این حوزه از تحقیق، دارای چالش‌هایی است و بخشی از فعالیت‌های علمی دانشمندان در این زمینه انجام می‌شود. در [۵،۶،۷،۸]، موضوع تشخیص گوینده و صحبت، به صورت جامع و با جزئیات بررسی شده است. در [۹]، با به کارگیری فرکانس‌های مدولاسیون زمانی-طیفی، یک روش جدید استخراج ویژگی برای مقاوم‌سازی سیستم‌های تشخیص گوینده ارایه شده است. به منظور بازیابی صحبت مبتنی بر محتوی، یک الگوریتم جدید مبتنی بر مدولاسیون چند مقیاسی زمانی-طیفی سیگنال صحبت در [۱۰] پیشنهاد شده است. در [۱۱]، مجموعه‌ای از روش‌های پنجره‌گذاری برای محاسبه‌ی MFCC[4] به منظور استفاده در تشخیص گوینده بحث شده است. در [۱۲]، یک سیستم تشخیص گوینده‌ی مبتنی بر شبکه‌های عصبی کانولوشنال پیشنهاد شده است که در آن از ترکیب مولفه‌های SE[5] و شبکه‌های عصبی کانولوشنال استفاده شده و SECNN [6] نام دارد. مشکل عدم تطابق[7] ایجاد شده به دلیل شرایط محیطی در [۱۳] بحث شده است و با به کارگیری روش‌های جداسازی سیگنال صحبت دوگوشی[8]، برای حل این مشکل راه‌حل‌هایی ارایه شده است. از آنجاییکه در اکثر کاربردهای واقعی، تشخیص گوینده برای طول سیگنال کم اهمیت بالایی دارد این موضوع در [۱۴] بررسی شده است و یک مجموعه بردار ویژگی جدید برای تشکیل مدل مخلوط گاوسی در شرایط طول پایین سیگنال پیشنهاد شده است. در [۱۵]، برای تشخیص هویت افراد از استخراج ویژگی‌های مرتبط و ایجاد یک مدل بر اساس این ویژگی‌ها استفاده شده است. در [۱۶]، با معرفی ویژگی‌های آکوستیکی طولانی مدت (LTA)[9] تلاش شده است که

[1] Text Dependent
[2] Text Independent
[3] Convolutional Neural Network
[4] Mel Frequency Cepstral Coefficients (MFCC)
[5] Squeeze-and-Excitation (SE)
[6] Squeeze-and-Excitation Convolutional Neural Network
[7] Mismatch
[8] Binaural
[9] Long-Term Acoustic

نمایش تُنُکی از سیگنال صحبت ارایه شود و بر اساس آن یک سیستم تشخیص گوینده مستقل از متن پیشنهاد شده است.

در [۱۷]، یک روش یادگیری انتقال بر پایه‌ی شبکه‌ی ResNet[۱] برای تشخیص گوینده در مخابرات رادیویی نظامی ارایه شده است. یک سیستم تشخیص گوینده‌ی اتوماتیک مبتنی بر یادگیری عمیق و با استفاده از تکنیک های CNN و LSTM[۲] در [۱۸] پیشنهاد شده است. در [۱۹] نیز از ترکیب CNN و LSTM برای تشخیص گوینده در EFPI[۳] استفاده شده است. در [۲۰]، با مدل سازی گوینده با استفاده از کوانتیزاسیون برداری و همینطور به‌کارگیری MFCC برای استخراج ویژگی، عملیات تشخیص گوینده انجام می‌شود. در [۲۱]، از مدل Res2Net برای تشخیص گوینده استفادده شده است. در [۲۲]، علاوه بر پیشنهاد یک مکانیزم خودتوجه جدید، یک مطالعه‌ی تجربی روی ترانسفورمرهای[۴] مختلف با و بدون مکانیزم پیشنهادی صورت گرفته است. همچنین در [۲۳]، سیستم‌های تشخیص گوینده مبتنی بر ترانسفورمر برای زبان قزاقی مطالعه شده است. در [۲۴]، نشان داده شده است که ترکیب مدل‌های پیش‌بینی غیرخطی با مدل کلاسیک LPCC[۵] منجر به بهبود دقت در سیستم‌های تشخیص گوینده می‌گردد. در [۲۵] یک تکنیک یادگیری عمیق تعبیه‌شده مقاوم برای تشخیص گوینده ارایه شده است که در آن یک روش خود-نظارتی مبتنی بر SKD[۶] به منظور بهره‌برداری از اطلاعات پنهان داده‌های بدون برچسب طراحی شده است. در [۲۶]، تاثیر سبک صحبت کردن و تنوع شرایط صوتی روی کارایی تشخیص گوینده بررسی شده است.

در [۲۷]، سه سیستم نگاشت گوینده i-vector، t-vector و x-vector برای موضوع تشخیص گوینده توسعه داده شده است. این تحقیق با در نظر گرفتن شرایط عدم تطابق که توسط NIST[۷] و SRE[۸] معرفی شده، انجام شده است. در [۲۸]، به منظور کاهش اثر نویزهای مختلف در دقت سیستم‌های تشخیص گوینده، از یک روش تخمین طیف مبتنی بر پنجره‌گذاری چند مخروطی استفاده شده است. در [۲۹]، عنوان شده است که روش‌های مبتنی بر شبکه های عصبی عمیق برای دستگاه‌های متحرک مناسب نیستند و برای حل این مشکل یک سیستم تشخیص گوینده مبتنی بر حاشیه‌ی افزودنی برای شبکه‌ی موبایل[۹] ارایه شده است. در [۳۰]، به منظور ارتقای کیفیت سیستم‌های تشخیص گوینده‌ی مبتنی بر یادگیری عمیق، روش AM-SincNet[۱۰] ارایه شده است که بر اساس مدل شبکه عصبی SincNet پایه‌ریزی شده است با این تفاوت که از یک لایه‌ی بهبودیافته AM-Softmax استفاده می‌کند. مدل مخلوط گاوسی[۱۱] قادر است به دقت مشابه گوش انسان در صحبت طولانی مدت برسد اما برای صحبت کوتاه مدت دچار خطا می‌شود [۳۱]؛ به منظور ارتقای دقت سیستم‌های مبتنی بر مدل مخلوط گاوسی در صحبت کوتاه مدت، در [۳۱] یک مدل جدید بر اساس آموزش یک شبکه ی عصبی کانولوشنال معرفی شده است. در [۳۲]، از نسبت واریانس درون-کلاسی و برون-کلاسی ویژگی‌ها استفاده شده است که این ویژگی‌ها شامل ضرایب LPCC و پیش‌بینی خطی ادراکی[۱۲] است.

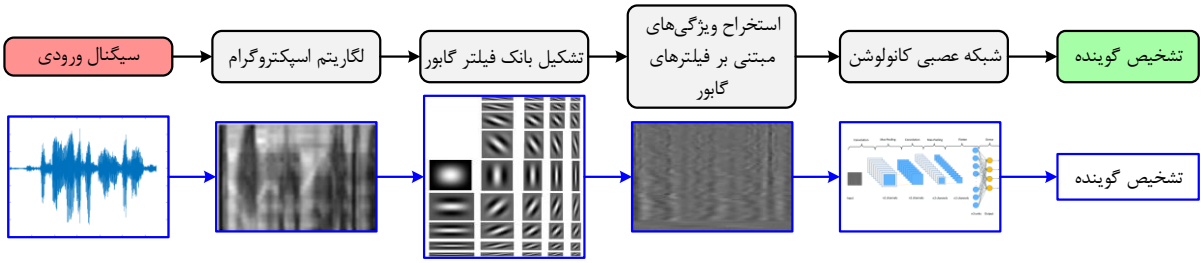
در [۳۳]، مدل مخلوط گاوسی برای تعیین توزیع احتمال صحبت خنثی[۱۳] یک گوینده استفاده شده است و بر اساس آن، یک مدل جدید برای مینیمم نمودن عدم تطابق بین صحبت خنثی و احساسی پیشنهاد شده است. در [۳۴]، یک مدل جدید مبتنی بر شبکه‌ی عصبی کانولوشنال با در نظر گرفتن عدم قطعیت با استفاده از نتروسافیک[۱۴] ارایه شده است. در [۳۵]، با ترکیب شبکه‌ی ResNet و مکانیزم خودتوجه[۱۵]، سعی شده است با تعداد پارامترهای کمتر و محاسبات کمتر، سیستم تشخیص گوینده به دقت بالاتری دست یابد. در [۳۶]، ابتدا MFCC برای استخراج ویژگی های سیگنال صحبت به کار گرفته می‌شود، سپس با استفاده از SOFM[۱۶]، بُعد ویژگی کاهش می‌یابد و در نهایت با استفاده از یک شبکه‌ی عصبی پرسپترون چندلایه، گوینده تشخیص داده می‌شود. در [۳۷]، برای تشخیص گوینده، از رویکردهای یادگیری مخالف مبتنی بر

[1] Residual Network
[2] Long Short-Term Memory
[3] external Fabry-Perot interferometric
[4] Transformer
[5] Linear Predictive Cepstral Coefficient (LPCC)
[6] Smoothed Knowledge Distillation
[7] National Institute of Standards and Technology
[8] Speaker Recognition Evaluation
[9] Additive Margin MobileNet1D
[10] Additive Margin-SincNet
[11] Gaussian Mixture Model (GMM)
[12] Perceptual Linear Prediction (PLP)
[13] Neutral Speech
[14] Neutrosophic
[15] Self-attention
[16] Self Organizing Feature Map

PLDA[1] استفاده شده است که به نظر می‌رسد به عنوان یک مدل متغیر پنهان برای بازسازی i-vector باشد. در [۳۸]، یک پلتفرم جدید مبتنی بر یادگیری ماشین برای تشخیص خودکار گوینده پیشنهاد شده است که از یک سنسور آکوستیکی پیزوالکتریک انعطاف‌پذیر بهره می‌گیرد. در [۳۹]، یک الگوریتم وزن‌دهی موثر در زمان محاسبه‌ی آمارگان Baum-Welch مربوط به مدل مخلوط گاوسی در i-vector ارایه شده است. در [۴۰]، به منظور بهبود دقت سیستم تشخیص گوینده، از ترکیب ویژگی‌های MFCC و LPC استفاده شده است. در [۴۱]، با ارایه‌ی یک شبکه‌ی عصبی کانولوشنال یک بُعدی بر پایه‌ی مدل دو بُعدی آن، تعداد پارامترها و همینطور حجم محاسبات به مقدار قابل توجهی پایین آمده است و روی FPGA پیاده‌سازی شده است. قابل ذکر است که دقت سیستم پیاده‌سازی شده فقط یک درصد افت دقت داشته است.

۳-روش پیشنهادی

سیستم تشخیص گوینده پیشنهادی مبتنی بر ویژگی‌های زمان-فرکانس گابور و شبکه‌های عصبی کانولوشن که فلوچارت آن در شکل (۱) آمده است در این بخش توضیح داده می‌شود. برای این منظور ابتدا دلایل استفاده از ویژگی-های گابور، سپس استخراج اسپکتروگرام و تشکیل فیلترهای گابور تشریح می‌شود و در نهایت چگونگی استخراج ویژگی‌های مبتنی بر بانک فیلتر گابور و مدل شبکه‌ی عصبی کانولوشن ارایه می‌شود.



شکل ۱- فلوچارت روش پیشنهادی.

۳-۱-دلایل استفاده از ویژگی های زمان-فرکانس گابور

بیشتر سیستم‌های تشخیص گوینده بر مبنای پردازش زمان-کوتاه[2] سیگنال‌های صوتی کار می کنند. پردازش زمان-کوتاه به این صورت است که با قطعه کردن سیگنال به قطعه‌های ۲۰ الی ۲۵ میلی ثانیه‌ای، و سپس استخراج ویژگی از این قطعه‌ها، بردار ویژگی تشکیل می‌شود. MFCC، PLP و مشتق‌های اول و دوم آنها نمونه‌ای از این نوع ویژگی‌ها هستند و چون در این موارد از بعد زمان برای استخراج ویژگی استفاده نمی‌شود باعث افزایش مقاومت سیستم‌های تشخیص گوینده در محیط‌های نویزی می‌شود. همچنین، نمایش زمان-فرکانس سیگنال‌های صوتی برای تمایز آواها مورد مطالعه قرار گرفته است که دارای قدرت تفکیک مناسبی است [۴۲]. وقتی سیگنال‌های صوتی در حوزه‌ی زمان-فرکانس نمایش داده می‌شوند، می‌توان از تبدیل‌ها و فیلترهای مختلف استفاده کرد تا بردار ویژگی از آنها استخراج شود. یکی از فیلترهایی که می‌تواند به حوزه‌ی زمان-فرکانس سیگنال‌های صوتی اعمال شود، فیلتر گابور است. مطالعات نشان می‌دهد فیلتر گابور برای تحلیل ویژگی‌های صوتی مناسب است. آواهایی که با نگه داشتن هوا در حنجره و سپس آزاد کردن ناگهانی هوا به وجود می‌آیند، آواهای سایشی که با اصطکاک تنفس تولید می‌شوند و نیز آواهایی که از بینی خارج شده و تولید می شوند، به راحتی با فیلترهای گابور قابل تفکیک هستند [۴۳]. استخراج ویژگی‌های گابور نیاز به طراحی مجموعه‌ای از فیلترهای گابور برای دریافت اطلاعات در حوزه‌ی زمان، فرکانس و الگوهای زمان-فرکانس دارد. فیلترهای گابور از فیزیولوژی صدای انسان الهام شده‌اند و بر روی تبدیل زمان-فرکانس سیگنال‌های صوتی اعمال می‌شوند. برای انتخاب مجموعه‌ی مناسبی از فیلترهای گابور، یک بانک فیلتر بهینه طراحی می‌شود تا با اعمال هر فیلتر از این بانک فیلتر، مجموعه‌ای از ویژگی‌ها استخراج شوند. ویژگی‌های صوتی استخراج شده از فیلترهای گابور منجر به افزایش دقت مدل‌های کلاس‌بندی برای کاربردهای تشخیص

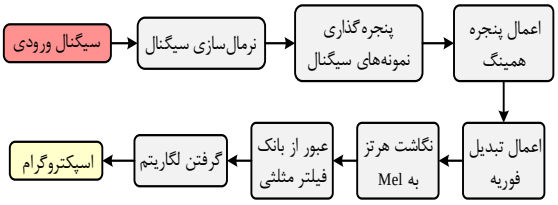
[1] Probabilistic Discriminant Analysis

[2] Short-time

گوینده در محیط‌های نویزی می‌شود [۱۰]. در این مقاله، مزایای ذکر شده، انگیزه‌ی اصلی برای استفاده از ویژگی‌های زمان-فرکانس مبتنی بر فیلترهای گابور است.

۳-۲-استخراج اسپکتروگرام

یکی از مراحل مهم در استخراج ویژگی‌های گابور، استخراج اسپکتروگرام از سیگنال صوتی است که مراحل استخراج اسپکتروگرام از یک سیگنال در فلوچارت شکل (۲) آمده است.



شکل ۲- فرآیند استخراج اسپکتروگرام از یک سیگنال.

ابتدا با استفاده از رابطه‌ی (۱)، توزیع نمونه‌های سیگنال به توزیع نرمال نزدیک می‌شود.

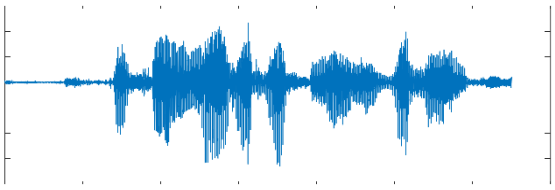
$$f(i)=\frac{f(i)-\mu}{\sigma} \qquad (۱)$$

برای استخراج اسپکتروگرام از سیگنال صوتی، پارامترهایی از قبیل طول پنجره (W)، شیفت (انتقال) پنجره (S)، حداقل و حداکثر فرکانس باید تنظیم شوند. برای اینکار، ابتدا یک قاب از ابتدای سیگنال با اندازه‌ی W جدا نموده سپس این قاب به اندازه‌ی S به سمت جلو حرکت داده می‌شود تا قاب دوم حاصل شود و این کار تا رسیدن به انتهای سیگنال ادامه می‌یابد.

پارامترهای W و S بر حسب میلی‌ثانیه هستند و با توجه به اینکه سیگنال نمونه‌برداری شده بر اساس تعداد نمونه‌ها است باید تعداد نمونه‌های هر قاب W میلی‌ثانیه‌ای تعیین شود. همینطور باید مشخص شود که برای رسیدن به قاب بعدی باید از چند نمونه عبور کنیم تا معادل S میلی‌ثانیه باشد. در هر سیگنال نمونه‌برداری شده، اگر زمان (بر حسب ثانیه) در نرخ نمونه‌برداری ضرب شود تعداد نمونه‌های مربوط به آن زمان بدست می‌آید. در روابط زیر پارامترهای Mو N به ترتیب تعداد نمونه‌های مربوط به طول پنجره‌ی W و انتقال S را نشان می‌دهند.

$$N=\frac{W\times f_s}{1000}\ ,\ M=\frac{S\times f_s}{1000} \qquad (۲)$$

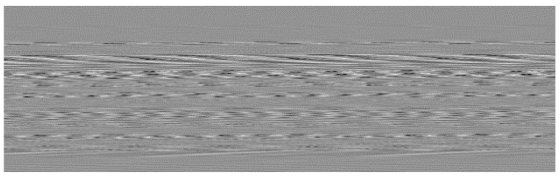
که در رابطه‌ی بالا f_s نرخ نمونه‌برداری است. به عنوان مثال سیگنال شکل (۳) را در نظر بگیرید.



شکل ۳- یک نمونه سیگنال صوتی با طول 2.95 ثانیه.

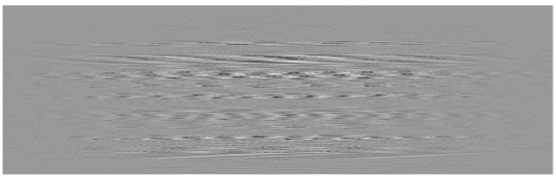
طول سیگنال شکل (۳) برابر 2.95s با نرخ نمونه‌برداری 44100 است در نتیجه تعداد 130095 نمونه از آن حاصل می‌شود. بنابراین، اگر W=25ms و S=10ms باشد آنگاه تعداد نمونه‌های هر پنجره N=1103 و تعداد نمونه‌ها برای انتقال M=441 است. برای اینکه مرکز اولین پنجره در ابتدای سیگنال و مرکز آخرین پنجره در انتهای سیگنال قرار گیرد، به اندازه نصف طول پنجره (550 نمونه) در ابتدا و انتهای سیگنال صفر قرار داده می‌شود.

مهمترین پارامتر در محاسبه تعداد کل پنجره‌های یک سیگنال S است. برای سیگنال مثال بالا، از آنجاییکه S=10ms و طول کل سیگنال 2.95s است 295 پنجره حاصل می‌شود که این تعداد از تقسیم طول سیگنال (بر اساس زمان یا بر اساس تعداد نمونه‌ها) بر طول انتقال بدست می‌آید (2.95/0.01=130095/441=295). بنابراین، برای سیگنال فوق 295 پنجره با تعداد 1103 نمونه در هر پنجره داریم که به صورت یک ماتریس 1103×295 ذخیره می‌شود که هر سطر از این ماتریس حاوی نمونه‌های یک پنجره است. شکل (۴) ماتریس مربوط به پنجره‌های یک نمونه سیگنال را نشان می‌دهد.



شکل ۴- پنجره‌گذاری نمونه‌های سیگنال.

در مرحله بعد، به منظور تاکید روی اطلاعات مرکزی پنجره، به تعداد پنجره‌های سیگنال، پنجره‌ی همینگ با طول هم اندازه با طول پنجره‌ی سیگنال در نظر گرفته شده و در ماتریس قبلی ضرب می‌شود. در حقیقت هر سطر ماتریس قبل در پنجره‌ی همینگ ضرب می‌شود. نتیجه‌ی حاصلضرب سیگنال پنجره‌گذاری شده در پنجره‌ی همینگ در شکل (۵) آمده است.



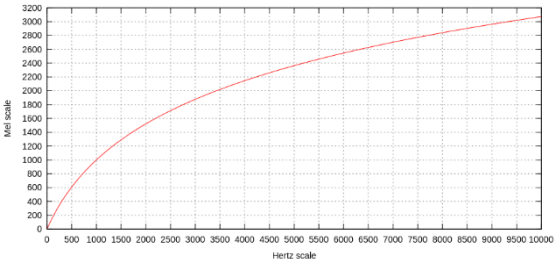
شکل ۵- ضرب پنجره‌های سیگنال در پنجره‌ی همینگ.

در ادامه، از هر پنجره تبدیل فوریه گرفته می‌شود تا مولفه‌های فرکانسی هر پنجره بدست آید. از آنجایی که تبدیل فوریه به نمونه‌های هر پنجره اعمال می‌شود، می‌توان تعداد مولفه‌های فرکانسی این تبدیل را برابر تعداد نمونه-های آن پنجره در نظر گرفت. در اینجا برای افزایش سرعت از تبدیل فوریه‌ی سریع استفاده شده است. برای استفاده از تبدیل فوریه‌ی سریع، باید تعداد مولفه‌های فرکانسی توانی از 2 باشد که در اینجا نزدیکترین توان 2 بزرگتر از تعداد نمونه‌ها به عنوان تعداد مولفه‌های فرکانسی در نظر گرفته می‌شود که از رابطه‌ی زیر بدست می‌آید.

$$NC = 2^{\lceil \log_2 N \rceil} \qquad (۳)$$

در رابطه‌ی بالا $\lceil\ \rceil$ حد بالای عدد صحیح و N تعداد نمونه های یک پنجره است. برای سیگنال مثال قبل NC=2048 بوده و در نتیجه، بعد از اعمال تبدیل فوریه یک ماتریس 2048×295 حاصل می‌شود که هر سطر آن مولفه‌های فرکانسی یک پنجره است.

بعد از محاسبه‌ی تبدیل فوریه‌ی هر پنجره، کل بازه‌ی فرکانسی این تبدیل، به چند بخش تقسیم می‌شود که هر بخش یک کانال نامیده می‌شود. هر کانال با یک مرکز مشخص می‌شود که نماینده‌ی همه‌ی فرکانس‌های موجود در آن کانال است. در اینجا بازه‌ی فرکانسی سیگنال صحبت بین 64 تا 4000 هرتز در نظر گرفته شده است که این دو مقدار حداقل و حداکثر مولفه‌های فرکانسی هستند که باید از هر پنجره استخراج شوند. بقیه‌ی فرکانس‌های خارج از این بازه حذف می‌شوند. برای سیستم تشخیص گوینده، 23 کانال در نظر گرفته شده است و بدین معنی است که بازه‌ی 64 تا 4000 به 23 کانال تقسیم شده و مرکز هر کانال به عنوان نماینده‌ی آن کانال تعیین می‌شود. برای افزایش کارایی، از مقیاس Mel استفاده می‌شود. این مقیاس یک تبدیل بر روی مقیاس هرتز است که برای فرکانس‌های کمتر از 800 هرتز خطی و برای فرکانس‌های بالاتر از 1 کیلوهرتز لگاریتمی است. تبدیل Mel در واقع یک نگاشت برای فرکانس‌های قابل درک توسط گوش انسان است که در شکل (۶) نشان داده شده است.



شکل ۶- نگاشت هرتز به Mel.

همانطور که از تبدیل هرتز به Mel مشخص است، نگاشت Mel باعث فشرده‌سازی بازه‌ی فرکانس می‌شود به گونه‌ای که بازه‌ی 0 تا 10000 هرتز را به بازه‌ی 0 تا 3200 هرتز نگاشت می‌کند. اکنون ابتدا بازه‌ی اولیه فرکانس یعنی [64 4000] به مقیاس Mel تبدیل می‌شود که بازه‌ی [98 2146] حاصل می‌گردد. در ادامه، بازه‌ی Mel به 25 کانال با طول مساوی تقسیم شده، سپس هر کدام از مراکز موجود در مقیاس Mel به مقیاس هرتز تبدیل می‌شود که بازه‌ی نتیجه در جدول ۱ آمده است.

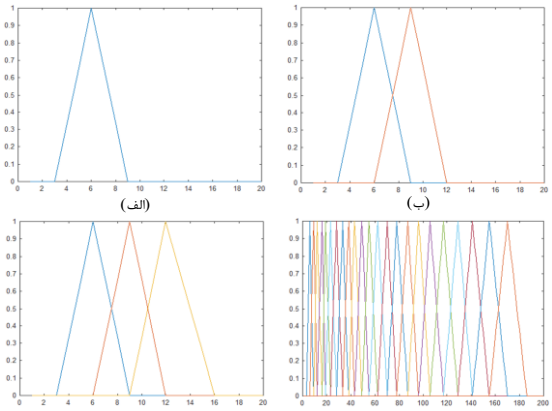
جدول ۱- بازه‌ی به دست آمده از تبدیل Mel، تبدیل Mel به هرتز و مراکز فرکانسی بر اساس نرخ نمونه‌برداری 44100.

کانال	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
بازه Mel	98	183	269	354	439	524	610	695	781	866	951	1037	1100	1207	1293	1378	1463	1549	1634	1719	1804	1890	1975	2060	2146
تبدیل Mel به هرتز	64	124	189	259	334	415	503	598	700	810	929	1057	1159	1344	1505	1678	1865	2067	2284	2519	2772	3045	3340	3657	4000
مراکز فرکانسی	3	6	9	12	16	19	23	28	33	38	43	49	55	62	70	78	87	96	106	117	129	141	155	170	186

علت در نظر گرفتن 25 کانال این است که در مراحل بعدی، بعد از پنجره‌گذاری کانال اول و آخر حذف شده و 23 کانال حاصل می‌شود که همان کانال‌های اصلی هستند. در ادامه مراکز فرکانسی فوق بر اساس نسبت آنها به نرخ نمونه‌برداری 44100 به بازه‌ی 0 تا 2048 که تعداد مولفه‌های فرکانسی در تبدیل فوریه بود، نگاشت می‌شوند که در نهایت بازه‌ی مشخص شده در جدول ۱ (سطر مربوط به مراکز فرکانسی) به دست می‌آید.

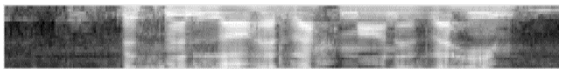
در ادامه، مراکز فرکانسی بالا با پنجره‌هایی به طول 3، پنجره‌گذاری می‌شوند و در هر مرحله، پنجره یک واحد به سمت جلو انتقال داده می‌شود. هر 3 مولفه‌ی متوالی در یک پنجره به عنوان نقطه‌ی شروع، وسط و پایان یک مثلث در

نظر گرفته می‌شود تا یک فیلتر مثلثی ایجاد شود. به عنوان مثال فیلتر مثلثی اول از نقطه‌ی 3 شروع شده و در نقطه‌ی 6 به پیک خود می‌رسد و در نقطه‌ی 9 به محل اول برمی گردد. اگر پنجره با طول 3 در همه‌ی مراکز فرکانسی لغزانده شود از 25 مرکز فرکانسی فوق 23 فیلتر مثلثی ایجاد می‌شود که در شکل (۷) چگونگی ایجاد این فیلترها نشان داده شده است.



شکل ۶- بانک فیلتر مثلثی، هر فیلتر مثلثی مربوط به یک کانال است.

هر کدام از فیلترهای فوق در یک بردار 2048 عنصری قرار داده می‌شود. به عنوان مثال وقتی فیلتر اول در این بردار قرار داده می‌شود عناصر 3 تا 9 همان مقادیر در فیلتر مثلثی بوده و مقادیر بقیه عناصر صفر می‌شود. علت این کار مشخص کردن مولفه‌های مهم در 2048 مولفه‌ی فرکانسی تبدیل فوریه است که در مراحل قبل، از هر پنجره اخذ شده است. با قرار دادن 23 فیلترمثلثی فوق در بردارهای 2048 عنصری، یک ماتریس 2048×23 حاصل می‌شود. اگر این ماتریس در ماتریس تبدیل فوریه‌ی پنجره‌ها با ابعاد 2048×295 ضرب شود این مفهوم را می‌رساند که برای هر پنجره، از 2048 مولفه‌ی فرکانسی در تبدیل فوریه، فقط آنهایی در نظر گرفته می‌شوند که در فیلتر مثلثی دارای مقادیر غیر صفر باشند. ذکر این نکته ضروری است که در اینجا هر کانال فرکانسی با یک فیلتر مثلثی در نظر گرفته شده است. بدین ترتیب با ضرب فیلتر مثلثی مربوط به هر کانال در هر پنجره، مولفه‌های فرکانسی آن کانال در آن پنجره حاصل می‌شود. بنابراین، در نهایت یک ماتریس 295×23 به دست می‌آید که مولفه‌های فرکانسی 23 کانال برای 295 پنجره است. این ماتریس 295×23 اسپکتروگرام سیگنال ورودی نامیده می‌شود که با اخذ لگاریتم از آن، اسپکتروگرام نهایی حاصل می‌گردد و در شکل (۸) نشان داده شده است.



شکل ۸- اسپکتروگرام نهایی سیگنال شکل ۲.

۳-۳- تشکیل فیلترهای گابور

بعد از استخراج اسپکتروگرام، باید فیلترهای گابور طراحی شوند. اولین گام در طراحی فیلترهای گابور تعین مرکز هر فیلتر است که از اهمیت بالایی برخوردار است. از آنجایی که فیلترهای گابور از نوع زمان-فرکانس هستند، مراکز آنها دارای دو بعد زمان و فرکانس هستند. در اینجا محاسبه مراکز توضیح داده می‌شود و برای راحتی کار، ابتدا مراکز زمان و سپس مراکز فرکانس محاسبه می‌شوند. ابتدا پارامتر c با استفاده از رابطه‌ی زیر به دست می‌آید:

$$C = d \times \frac{8}{v_k} \qquad (۴)$$

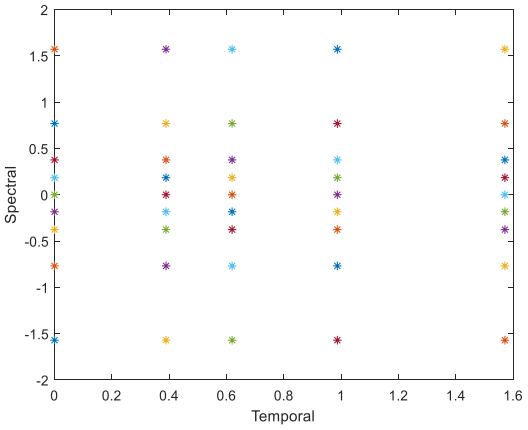
که در رابطه‌ی بالا d فاصله‌ی بین فیلترها را تنظیم نموده و v_k میزان پوشش فیلترها را کنترل می‌کند. به عنوان مثال اگر d=0.2 و v_k=3.5 باشد C=0.4571 خواهد بود. اکنون باید یک فاکتور برای پوشش فضای زمان تعیین شود. پوشش به این معناست که اگر یک مرکز زمان اولیه به صورت تکراری در یک فاکتور ضرب شود تا مراکز دیگر حاصل شوند، مراکز حاصل از این ضرب، چه فضایی از زمان را پوشش می‌دهند. فاکتور پوشش فضای زمان بصورت زیر محاسبه می‌شود.

$$spanK = \frac{1+\frac{C}{2}}{1-\frac{C}{2}} \qquad (۵)$$

برای مقدار C بالا، فاکتور پوشش $spanK$=1.59 حاصل می‌شود و بدین معناست که با ضرب توان‌های متوالی 1.59، مراکز فیلترها در حوزه‌ی زمان تعین می‌شوند. اگر حداکثر مرکز زمان $\pi/2$ که معادل با مقدار 1.57 است در نظر گرفته شود باید بصورت متوالی بر فاکتور پوشش تقسیم گردد تا به مرکز نزدیک شویم. با حرکت به سمت مرکز با استفاده از فاکتور پوشش، مقادیر 1.57، 0.98، 0.61 و 0.38 بدست می‌آید. مقدار 0 هم در کنار آنها قرار داده می‌شود تا فضای زمان از $\pi/2$ تا 0 پوشش داده شود. این مقادیر از رابطه زیر حاصل شده‌اند:

$$\omega_n^{i+1} = \omega_n^i \times spanK = \omega_n^i \times \frac{1+\frac{C}{2}}{1-\frac{C}{2}} \quad (۶)$$

که در رابطه‌ی بالا ω_n^{i+1} مرکز زمان در مرحله‌ی i+1 و ω_n^i مرکز زمان در مرحله‌ی i است. با روابطی مشابه روابط فوق، مراکز فرکانسی نیز محاسبه می‌شوند. با این تفاوت که فاکتور پوشش در بعد فرکانس، به مقدار v_k=2.04 تنظیم می‌شود و با شروع از اولین مرکز فرکانسی $\pi/2$ که 1.57 است به مراکز بعدی 0.76، 0.38 و 0.18 می‌رسیم و در نهایت به 0 می‌رسیم. همین مراکز در قست منفی محورها نیز ادامه داده می‌شوند. نتیجه‌ی عملیات تعیین مراکز در شکل (۹) آمده است.



شکل ۹- مراکز فیلترهای گابور در حوزه زمان و فرکانس.

مقادیر حاصل از این رابطه برای مراکز نشان می‌دهد که هر چه از صفر دور شویم فاصله‌ی مراکز بیشتر می‌شود. به عنوان مثال، از مرکز 0 با گام 0.38 به دومین مرکز می‌رسیم در حالی که از مرکز 0.98 با گام 0.59 به مرکز 1.57 می‌رسیم. بنابراین، هرچه از صفر دور می‌شویم فاصله‌ی بیشتری بین فیلترها قرار داده می‌شود. علت این کار این است که با این روش، در حوالی صفر فیلترهای بیشتری با ابعاد کوچکتر قرار می‌گیرد تا بتوانیم اطلاعات بیشتری در حوالی محورهای مختصات استخراج کنیم.

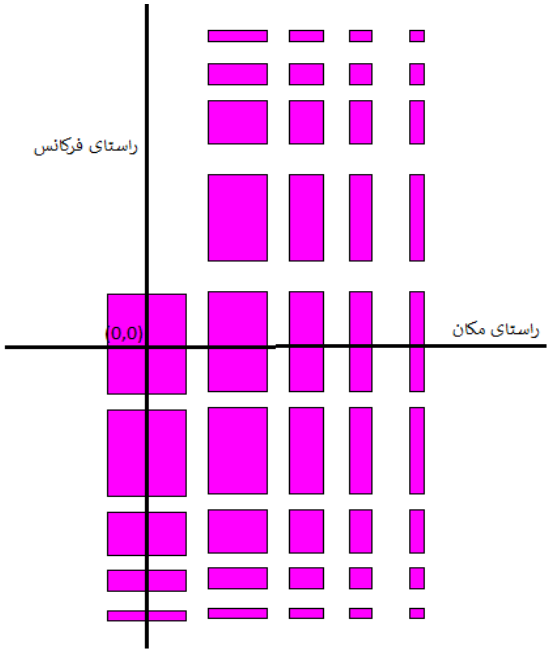
در بعد زمان 5 مرکز و در بعد فرکانس 9 مرکز حاصل شده است که با ترکیب‌های مختلف از این دو مجموعه 45 مرکز در صفحه‌ی زمان-فرکانس بدست می‌آید. مراکزی که بعد زمان آنها صفر و بعد فرکانس آنها منفی است را حذف می‌کنیم که در صفحه‌ی زمان-فرکانس، 4 مرکز این شرایط را دارند که باید از 45 مرکز بدست آمده حذف شوند. بنابراین، در نهایت 41 مرکز خواهیم داشت.

بعد از محاسبه‌ی مراکز فیلترها، باید طول و عرض هر فیلتر تعیین شوند. طول فیلتر در بعد زمان و فرکانس با روابط زیر محاسبه می‌شوند:

$$w_n = \frac{2\pi}{|\omega_n|} \times \frac{\upsilon}{2}, \quad w_k = \frac{2\pi}{|\omega_k|} \times \frac{\upsilon}{2} \quad (۷)$$

که در روابط بالا، w_n و w_k به ترتیب طول فیلتر در بعد زمان و بعد فرکانس هستند. υ یک ثابت است که در اینجا 3.5 در نظر گرفته می‌شود. حداکثر طول فیلتر در بعد زمان 40 و در بعد فرکانس 63 در نظر گرفته می‌شود. اگر طول فیلتر بدست آمده از روابط فوق، بزرگتر از مقدار حداکثر باشد طول حداکثر در نظر گرفته می‌شود. برای هر دو بعد زمان و فرکانس، هر چه اندازه‌ی مرکز بیشتر باشد طول فیلتر کمتر خواهد شد. در روابط بالا اگر مرکز یک فیلتر صفر باشد مخرج کسر صفر شده که باعث می‌شود طول پنجره بینهایت گردد. در این حالت برای آن بعد (زمان یا فرکانس)، حداکثر طول فیلتر در نظر گرفته می‌شود. بنابراین، بزرگترین فیلتر از نظر طول در راستای زمان و فرکانس، فیلتری است که مرکز آن (0,0) باشد. در محوری که مرکز فرکانسی فیلتر صفر باشد (محور افقی)، هر چه در راستای زمان از مرکز دور شویم (چه در جهت مثبت و چه در جهت منفی) و اندازه‌ی مرکز زمانی فیلتر را افزایش دهیم طول فیلتر در راستای فرکانس ثابت مانده و در راستای زمان کوچک‌تر می‌شود. با استدلالی مشابه، در محوری که مرکز زمانی فیلتر صفر باشد (محور عمودی)، هر چه در راستای فرکانس از مرکز دور شویم (چه در جهت مثبت و چه در جهت منفی) و اندازه مرکز فرکانسی فیلتر را افزایش دهیم طول فیلتر در راستای زمان ثابت مانده و در راستای فرکانس کوچک‌تر می‌گردد.

اگر در راستای خط $y=x$ حرکت کنیم باعث افزایش همزمان اندازه‌ی مراکز زمانی و فرکانسی فیلترها می‌شود. این کار باعث کاهش طول فیلترها در راستای زمان و فرکانس می‌گردد. با حرکت در راستای خط $y=-x$ نیز اتفاقی مشابه می‌افتد. با توجه به این استدلال ها، ابعاد فیلترهای حاصل از رابطه های فوق بصورت شماتیک در شکل (۱۰) نشان داده شده است.

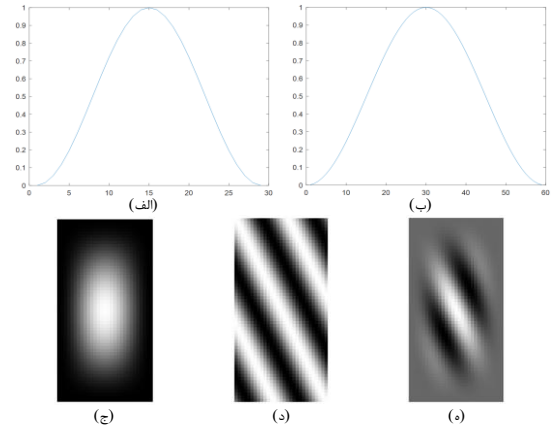


شکل ۱۰- ابعاد فیلترهای گابور.

بعد از تعیین ابعاد فیلترها، باید مقدار فیلتر تعیین شود. برای تعیین محتوای هر فیلتر، ابتدا با استفاده از رابطه‌ی زیر که رفتاری شبیه به یک پنجره‌ی هنینگ دارد، دو پنجره‌ی هنینگ یک بعدی در راستای زمان و فرکانس ایجاد می‌شود.

$$h(x) = 0.5 - 0.5 \times \cos(2\pi x) \qquad (۸)$$

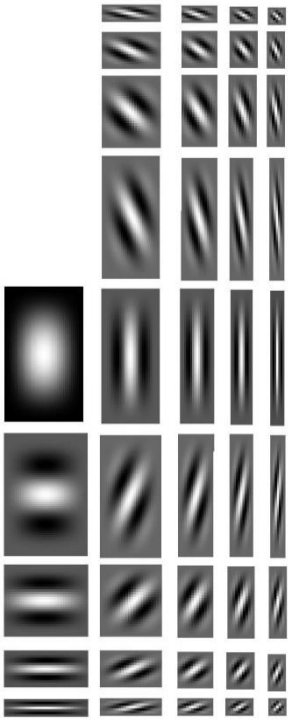
به عنوان مثال در فیلتر با ابعاد 59×29، یک پنجره‌ی هنینگ یک بعدی با 29 عنصر در راستای زمان مطابق شکل (۱۱- الف) و یک پنجره‌ی 59 عنصری در راستای فرکانس مطابق شکل (۱۱- ب) ایجاد می‌شود. سپس دو پنجره‌ی هنینگ فوق، که یکی در راستای افقی (زمان) و دیگری در راستای عمودی (فرکانس) است در هم ضرب شوند تا مطابق شکل (۱۱- ج)، فیلتر حاصل شود.



شکل ۱۱- مراحل تشکیل فیلتر گابور به صورت گرافیکی.

در مرحله بعد یک ماتریس با ابعاد 59×29 که برابر ابعاد فیلتر است، ایجاد می‌شود به گونه‌ای که ستون اول آن 1، ستون دوم آن 2، و در نهایت ستون 29ام آن 29 است. سپس یک ماتریس دیگر برای بعد فرکانس با ابعاد 59×29 ایجاد می‌گردد به گونه‌ای که سطر اول آن 1، سطر دوم آن 2، و سطر 59ام آن 59 است. اگر این دو ماتریس در هم ضرب شوند یک سیگنال دوبعدی سینوسی ایجاد می‌شود که در شکل (۱۱- د) نشان داده شده است. در مرحله آخر ماتریس مرحله قبل یعنی شکل (۱۱- ج) که از ضرب دو پنجره هنینگ حاصل شده بود در این ماتریس یعنی شکل (۱۱- د) ضرب می‌شود تا فیلتر نهایی نشان داده شده در شکل (۱۱- ه) حاصل شود.

اگر برای 41 فیلتر گابور، مراحل فوق تکرار شود فیلترهای با ابعاد مختلف در جهت‌های متفاوت ایجاد می‌شود. جهت سینوسی توسط ابعاد دو ماتریس بعد زمان و فرکانس تعیین می‌گردد. مجموعه‌ی این فیلترها به عنوان بانک فیلتر گابور در نظر گرفته می‌شوند که در شکل (۱۲) نشان داده شده‌اند.



شکل ۱۲- بانک فیلتر گابور.

۳-۴- استخراج ویژگی‌های مبتنی بر بانک فیلتر گابور

در نهایت برای استخراج ویژگی‌های گابور، فیلترهای گابور با اسپکتروگرام سیگنال کانوالو می‌شود که در این بخش به جزئیات آن پرداخته می‌شود. همانطور که قبلا اشاره شد،

23 کانال برای اسپکتروگرام در نظر گرفته شده است که اگر یک سیگنال دارای n قاب باشد، اسپکتروگرام آن یک ماتریس $n\times23$ است. به عنوان مثال برای سیگنال شکل (۳) که دارای 295 قاب بود ابعاد ماتریس اسپکتروگرام آن 295×23 است. به منظور کانوالو اسپکتروگرام با یک فیلتر گابور، ابتدا حاشیه‌ی ماتریس اسپکتروگرام صفرگذاری می‌شود. تعداد صفرهایی که به سمت چپ و راست اسپکتروگرام اضافه می‌شود نصف طول فیلتر گابور در راستای افقی (زمان) و تعداد صفرهایی که بالا و پایین ماتریس اسپکتروگرام اضافه می‌شود نصف طول فیلتر گابور در راستای عمودی (فرکانس) است. این عمل باعث می‌شود ابعاد ماتریس حاصل، 295×23 باشد که همان ابعاد اسپکتروگرام است. این ماتریس ویژگی‌های نهایی گابور می باشند.

اگر برای 41 فیلتر تعریف‌شده در شکل (۱۲) این عمل انجام شود 41 ماتریس ویژگی با ابعاد 295×23 حاصل می‌شود که با کنار هم قرار دادن این ماتریس‌ها یک ماتریس 295×943 به دست می‌آید که هر ستون آن یک بردار ویژگی متناظر با یک قاب می باشد. ابعاد 41 فیلتر گابور، متناظر با شکل (۱۲)، در جدول ۲ آمده است.

جدول ۲: ابعاد فیلترهای گابور.

	7×29	7×17	7×11	7×7
	15×29	15×17	15×11	15×7
	29×29	29×17	29×11	29×7
	59×29	59×17	59×11	59×7
69×39	69×29	69×17	69×11	69×7
59×39	59×29	59×17	59×11	59×7
29×39	29×29	29×17	29×11	29×7
15×39	15×29	15×17	15×11	15×7
7×39	7×29	7×17	7×11	7×7

همانطور که از جدول ۲ مشخص است، طول در راستای فرکانس (عمودی) برای همه‌ی فیلترهای گابور در سطر اول 7 است و برای سطرهای دوم، سوم، چهارم و پنجم به ترتیب 15، 29، 59 و 69 هستند. کانولوشن هر کدام از فیلترهای فوق با ماتریس اسپکتروگرام منجر به ایجاد یک ماتریس ویژگی با ابعادی مشابه با ابعاد ماتریس اسپکتروگرام می‌شود که علت این امر بزرگتر بودن ابعاد ماتریس اسپکتروگرام در مقایسه با ابعاد فیلترها است. به عنوان مثال، اگر فیلتر گابور 7×29 با ماتریس اسپکتروگرام با ابعاد 295×23 کانوالو شود ابعاد ماتریس ویژگی 295×23 خواهد بود و اگر فیلتر با ابعاد 15×29 با ماتریس اسپکتروگرام کانوالو شود باز هم ابعاد ماتریس ویژگی 295×23 می‌شود. تفاوت از اینجا ناشی می شود که در کانولوشن فیلتر گابور 7×29، با ماتریس اسپکتروگرام 295×23، هر بار 7 کانال (سطر) از ماتریس اسپکتروگرام در 7 سطر از فیلتر گابور ضرب شده تا یک ویژگی گابور را تولید کنند. همچنین، در کانولوشن فیلتر گابور 15×29، با ماتریس اسپکتروگرام 295×23، هر بار 15 کانال (سطر) از ماتریس اسپکتروگرام در 15 سطر از فیلتر گابور ضرب شده تا یک ویژگی گابور را تولید کنند. در حقیقت، در عمل کانولوشن، فیلتر گابور به صورت یک پنجره بر روی اسپکتروگرام قرار گرفته و روی کل سطح اسپکتروگرام لغزانده می‌شود.

از مثال فوق نتیجه می‌گیریم که تعداد کانال‌هایی از اسپکتروگرام که برای محاسبه یک ویژگی گابور مورد استفاده قرار می‌گیرند به بعد فرکانسی فیلتر گابور بستگی دارد. هرچه تعداد کانال‌های استفاده شده از اسپکتروگرام برای محاسبه یک ویژگی گابور بیشتر باشد، همبستگی بین ویژگی‌های گابور استخراج شده بیشتر خواهد شد. برای استدلال این نکته، این مثال را در نظر بگیرید: وقتی فیلتر گابور 7×29 بر روی ماتریس اسپکتروگرام قرار می‌گیرد تا با ضرب نقطه به نقطه‌ی این فیلتر با ماتریس اسپکتروگرام اولین ویژگی گابور را محاسبه کند، 7 کانال (سطر) از ماتریس اسپکتروگرام زیر این فیلتر قرار می‌گیرد و در محاسبه‌ی این ویژگی نقش دارند. این نشان می‌دهد که داده‌های سطر 7 ماتریس اسپکتروگرام، در محاسبه ویژگی سطر اول نقش دارند. ولی داده‌های سطر 7 به بعد (8، 9، 10 و ...) هیچ نقشی ندارند.

با استدلالی مشابه، برای محاسبه‌ی ویژگی‌های سطر 7، داده‌های سطر اول ماتریس اسپکتروگرام نقش دارند. بنابراین، ویژگی‌های با فاصله‌ی کمتر از 7 کانال (سطر) با هم همبستگی دارند. اگر ابعاد فیلتر گابور 15×29 باشد، واضح است که ویژگی‌های با فاصله‌ی کمتر از 15 کانال با هم همبستگی دارند که در مقایسه با حالت قبل همبستگی بیشتری ایجاد شده است. وجود همبستگی بالا بین ویژگی‌ها باعث کاهش کارایی در سیستم تشخیص گوینده و تشخیص صوت می‌شود. راه‌حل این مشکل این است که در جاهایی که همبستگی بین کانال‌ها زیاد است از بین همه‌ی کانال‌ها یک زیرمجموعه از آنها انتخاب گردد.

بر اساس استدلال ارائه شده، هر چه بعد فرکانسی فیلتر گابور بزرگتر باشد همبستگی بین کانال‌های متوالی

(سطرهای متوالی) در ماتریس ویژگی گابور بیشتر خواهد شد. بنابراین، فیلترهایی که بر روی محور افقی قرار دارند و دارای بزگترین طول با مقدار 69 در راستای فرکانس (عمود) هستند، بیشترین همبستگی بین کانال‌های ماتریس ویژگی را ایجاد می‌کنند. اگر فیلتر 69×29 بر روی ماتریس اسپکتروگرام 295×23 قرار گیرد، کل ماتریس در راستای فرکانس (عمود) در زیر فیلتر قرار خواهد گرفت چون بعد فرکانسی فیلتر 69 و بعد فرکانسی اسپکتروگرام 23 است که در این راستا فیلتر خیلی بزرگتر از ماتریس اسپکتروگرام است.

این نشان می دهد که وابستگی شدید بین تمام کانال‌ها در ماتریس ویژگی وجود دارد و برای حذف این وابستگی فقط یک کانال از بردار ویژگی باید انتخاب گردد که باید کانال با بیشترین اطلاعات انتخاب شود. اگر فیلتر بزرگتر از سیگنال باشد سیگنال باید صفرگذاری شود. بهترین حالت زمانی است که فیلتر در مرکز اسپکتروگرام قرار می‌گیرد. در این حالت کمترین تعداد صفر و بیشترین اطلاعات از سیگنال در زیر فیلتر قرار می‌گیرد. بنابراین، کانال مرکزی ماتریس ویژگی، دارای بیشترین اطلاعات است چون از کانولوشن فیلتر گابور با قرار گرفتن در مرکز اسپکتروگرام حاصل شده است. این کانال همان کانال انتخابی خواهد بود. بنابراین، اگر فیلتر گابور 69×29 باشد چون همه‌ی کانال‌های بردار ویژگی با هم همبستگی شدید دارند، فقط کانال مرکزی (در 23 کانال) یعنی کانال شماره‌ی 12 نگهداری شده و بقیه‌ی کانال‌ها حذف می‌شوند. در این حالت، ماتریس ویژگی گابور با ابعاد 295×23 به یک ماتریس با ابعاد 295×1 تبدیل می‌گردد. این حالت سخت‌ترین حالت کاهش است که در آن از بین 23 کانال موجود، فقط 1 کانال انتخاب شده است. برای همه فیلترهای گابور روی محور افقی این انتخاب انجام می‌شود چون بعد فرکانسی همه‌ی این فیلترها 69 است.

اگر بعد فرکانس فیلتر گابور کاهش یابد همبستگی کانال‌ها در ماتریس ویژگی کمتر می‌شود و باید به جای یک کانال، کانال‌های بیشتری انتخاب شود. برای تعیین تعداد کانال‌های انتخابی از بین 23 کانال، بر اساس طول فرکانسی فیلتر گابور معیاری تعیین می‌گردد. برای این کار ابتدا طول فرکانسی فیلتر گابور بر 4 تقسیم می‌شود که یک عدد بدست می‌آید. حد پایین این عدد مشخص کننده‌ی گام حرکت از کانال مرکزی 12 به سمت آخرین کانال یعنی 23 و اولین کانال یعنی 1 است. به عنوان مثال اگر ابعاد فیلتر گابور 15×29 باشد بعد فرکانسی 15 بر 4 تقسیم شده و حد پایین بدست می‌آید که مقدار آن 3 است. حالا از کانال مرکزی 12 با گام 3 به کانال آخر حرکت می‌کنیم که به کانال‌های 15، 18 و 21 می‌رسیم و سپس از کانال مرکزی 12 با گام 3 به سمت کانال اول حرکت می‌کنیم که به کانال های 9، 6 و 3 می‌رسیم. بنابراین برای همه فیلترهای گابور که بعد فرکانس آنها 15 است، کانال‌های انتخابی 7 کانال با شماره‌های 3، 6، 9، 12، 15، 18 و 21 هستند و ماتریس ویژگی گابور برای همه‌ی این فیلترها 295×7 است.

برای فیلتر گابور با بعد فرکانسی 7، چون گام 1 بدست می‌آید، همه‌ی کانال‌های ماتریس ویژگی باید انتخاب شوند و ماتریس ویژگی گابور برای همه‌ی این فیلترها 295×23 بوده که با ابعاد ماتریس اسپکتروگرام برابر است. ابعاد ماتریس‌های ویژگی برای همه‌ی فیلترهای گابور در جدول ۳ نشان داده شده است. در ماتریس ویژگی نهایی، دو قاب از اول و ۲ قاب از آخر سیگنال حذف می‌گردد چون این قاب‌ها بیشتر حاوی داده‌های صفر بوده و اطلاعات گوینده را در بر ندارند. بنابراین، تعداد قاب‌های نهایی 291 است.

جدول ۳: ابعاد ماتریس‌های ویژگی برای همه‌ی فیلترهای گابور.

	23×291	23×291	23×291	23×291
	7×291	7×291	7×291	7×291
	3×291	3×291	3×291	3×291
	1×291	1×291	1×291	1×291
1×291	1×291	1×291	1×291	1×291
1×291	1×291	1×291	1×291	1×291
3×291	3×291	3×291	3×291	3×291
7×291	7×291	7×291	7×291	7×291
23×291	23×291	23×291	23×291	23×291

برای سیگنال شکل (۳)، اگر همه‌ی ماتریس‌های ویژگی گابور کنار هم قرار بگیرند یک ماتریس 291×311 حاصل می‌شود که عدد 291 تعداد قاب‌های این سیگنال و 311 تعداد بردارهای ویژگی است. هر ستون در ماتریس بردار ویژگی مربوط به یک قاب از سیگنال است که به عنوان یک نمونه‌ی مستقل در کلاس‌بندی سیگنال برای کاربردهای تشخیص گوینده مورد استفاده قرار می‌گیرد.

۳-۵- معماری شبکه عصبی کانولوشن پیشنهادی

در سیستم تشخیص هویت گوینده، بعد از استخراج ویژگی از سیگنال‌های صوتی یک کلاس‌بند انتخاب می‌شود تا سیگنال‌های صوتی بر اساس هویت آنها تشخیص داده

شوند. از آنجاییکه اخیراً شبکه‌های عصبی کانولوشنال در بسیاری از کاربردها منجر به نتایج خوبی شده است [۴۴،۴۵،۴۶]، اینجا از شبکه‌های عصبی کانولوشن استفاده شده است که یکی از روش‌های مبتنی بر یادگیری عمیق است. جدول ۴ مشخصات شبکه‌ی عصبی کانولوشن پیشنهادی را نشان می‌دهد.

از آنجایی که داده‌های مورد نیاز برای سیستم تشخیص هویت گوینده بردارهای ویژگی گابور با ابعاد 300 است، هر ورودی شبکه باید قابلیت دریافت یک نمونه را داشته باشد. بنابراین، لایه‌ی ورودی در شبکه یک بردار با 300 عنصر قرار داده می‌شود. ساختار و ابعاد لایه‌های شبکه در جدول زیر آورده شده است.

جدول ۴- ساختار شبکه عصبی کانولوشن پیشنهادی.

عنوان لایه	نوع لایه	ابعاد ورودی	ابعاد فیلتر	تعداد فیلتر	طول گام	ابعاد خروجی
ورودی	ورودی	1×300×1	-	-	-	1×300×1
لایه اول	کانولوشن	1×300×1	1×3×1	8	1×3	1×100×8
	نرمال سازی	-	-	-	-	-
	Relu	-	-	-	-	-
لایه دوم	کانولوشن	1×100×8	1×5×8	16	1×5	1×20×16
	نرمال سازی	-	-	-	-	-
	Relu	-	-	-	-	-
لایه سوم	کانولوشن	1×20×16	1×4×16	32	1×4	1×5×32
	نرمال سازی	-	-	-	-	-
	Relu	-	-	-	-	-
لایه سوم	کانولوشن	1×5×32	1×5×32	5	1×5	1×1×5
	نرمال سازی	-	-	-	-	-
	Relu	-	-	-	-	-
تماما متصل	تماما متصل	1×1×20	-	-	-	20

۴- پایگاه‌های داده و تنظیم پارامترها

در این بخش ابتدا پایگاه‌های داده‌ی TIMIT و Aurora2 تشریح شده سپس به مقدار پارامترها و چگونگی تعیین آنها پرداخته می‌شود. روش پیشنهادی با زبان برنامه‌نویسی متلب پیاده‌سازی شده و بر روی یک ماشین با پردازنده 2.26 GHz Corei7 و حافظه 6GB اجرا شده است. برای استخراج ویژگی‌های گابور از کتابخانه پایه گابور نسخه دوم استفاده شده است.[1]

۴-۱- پایگاه داده‌ی TIMIT

پایگاه داده TIMIT یکی از رایج‌ترین پایگاه‌های داده‌ی صوتی است و برای بهبود و ارزیابی سیستم‌های پردازش صوت در سال ۱۹۹۰ طراحی شده است [۴۷]. این پایگاه داده شامل 630 گوینده از 8 منطقه آمریکا با لهجه‌های مختلف است که هر گوینده 10 جمله را ادا می‌کند و در مجموع شامل 6300 جمله‌ی ادا شده توسط کل گویندگان است. جدول ۵ تعداد گویندگان هر منطقه را بر حسب جنس آنها نشان می‌دهد.

جدول ۵- توزیع تعداد گویندگان هر منطقه.

منطقه	مرد	زن	جمع
۱	۳۱ (۶۳٪)	۱۸ (۲۷٪)	۴۹ (۸٪)
۲	۷۱ (۷۰٪)	۳۱ (۳۰٪)	۱۰۲ (۱۶٪)
۳	۷۹ (۶۷٪)	۲۳ (۲۳٪)	۱۰۲ (۱۶٪)
۴	۶۹ (۶۹٪)	۳۱ (۳۱٪)	۱۰۰ (۱۶٪)
۵	۶۲ (۶۳٪)	۳۶ (۳۷٪)	۹۸ (۱۶٪)
۶	۳۰ (۶۵٪)	۱۶ (۳۵٪)	۴۶ (۷٪)
۷	۷۴ (۷۴٪)	۲۶ (۲۶٪)	۱۰۰ (۱۶٪)
۸	۲۲ (۶۷٪)	۱۱ (۳۳٪)	۳۳ (۵٪)
مجموع	۴۳۸ (۷۰٪)	۱۹۲ (۳۰٪)	۶۳۰ (۱۰۰٪)

جملاتی که توسط گویندگان ادا می‌شود با علائم SA، SX و SI برچسب خورده‌اند. که هر گوینده 5 جمله‌ی SX، 3 جمله‌ی SI و 2 جمله‌ی SA را ادا می‌کند. میانگین طول هر جمله 3 ثانیه است و تمام جملات در محیط بدون نویز با فرکانس نمونه‌برداری 16kHz ضبط شده‌اند. این پایگاه داده به دو زیر مجموعه آموزش و تست تقسیم می‌شود که معمولا بین 20 تا 30 درصد داده‌ها به عنوان داده‌ی تست و 70 تا 80 درصد داده‌ها به عنوان داده‌ی آموزش مورد استفاده قرار می‌گیرند. تقسیم‌بندی باید به گونه‌ای باشد که نباید هیچ گوینده‌ای در هر دو زیرمجموعه‌ی آموزش و تست مورد استفاده قرار گیرد و باید از تمام 8 ناحیه در هر دو زیر مجموعه آموزش و تست مورد استفاده قرار گیرد.

۴-۲- پایگاه داده‌ی Aurora2

پایگاه داده‌ی دوم که در اینجا استفاده شده است، Aurora2 می‌باشد. این مجموعه داده حاوی گفتار متصل متشکل از ارقام انگلیسی با نرخ نمونه برداری 44.1kHz است [۴۸]. در پایگاه داده‌ی Aurora2 دو مجموعه‌ی

[1] Gabor Filter Bank Feature Extraction v2.0 - Author: Marc René Schädler, Email :marc.r.schaedler@uni-oldenburg.de Institute: Medical Physics / Carl-von-Ossietzky University Oldenburg, Germany

مجزای آموزشی و تست مشخص شده است. داده‌های آموزش به دو دسته تقسیم می‌شوند: داده‌های تمیز و داده‌های نویزی. داده‌های تست نیز به سه دسته A، B و C تقسیم می‌شوند که هر یک هم شامل گفتار تمیز و هم شامل گفتار نویزی با نویزهای مشخص شده برای هر مجموعه در نسبت‌های سیگنال به نویز5، 10، 15 و 20 دسی بل می‌باشد.

مجموعه گفتار نویزی A شامل نویزهای جمع‌پذیر مترو[۱]، ماشین[۲]، همهمه[۳] و نمایشگاه[۴] به همراه فیلتر کانال G.712 است. مجموعه گفتار نویزی B شامل نویزهای جمع‌پذیر رستوران[۵]، خیابان[۶]، فرودگاه[۷] و ایستگاه قطار[۸] به همراه فیلتر کانال G.712 می‌باشد. مجموعه گفتار نویزی C شامل نویزهای جمع‌پذیر مترو و خیابان است که هر یک با نویزهای مجموعه A و B مشترک هستند. اما در مجموعه‌ی C از فیلتر کانال MIRS استفاده شده، که متفاوت از نویز کانال استفاده شده در مجموعه‌ی A و B است. مشخصات پایگاه داده Aurora2 در جدول ۶ آمده است.

جدول ۶- مشخصات پایگاه داده Aurora2 [۴۹].

واژگان بیان‌شده	دنباله‌های عددی پیوسته (۰-۹)	
نمونه‌برداری	44.1kHz, 16 bits, mono	
مرد	۱۱۱ گوینده	۲۱-۷۰ سال
زن	۱۱۴ گوینده	۱۷-۵۹ سال
آموزش	۸۴۴۰ جمله	شرایط مختلف
	Subway, Babble, Car, Exhibition hall	
	20dB, 15dB, 10dB, 5dB and clean	
تست	۱۰۰۱ جمله	
	Subway, Babble, Car, Exhibition hall	
	10dB, 5dB 0dB, -5dB	

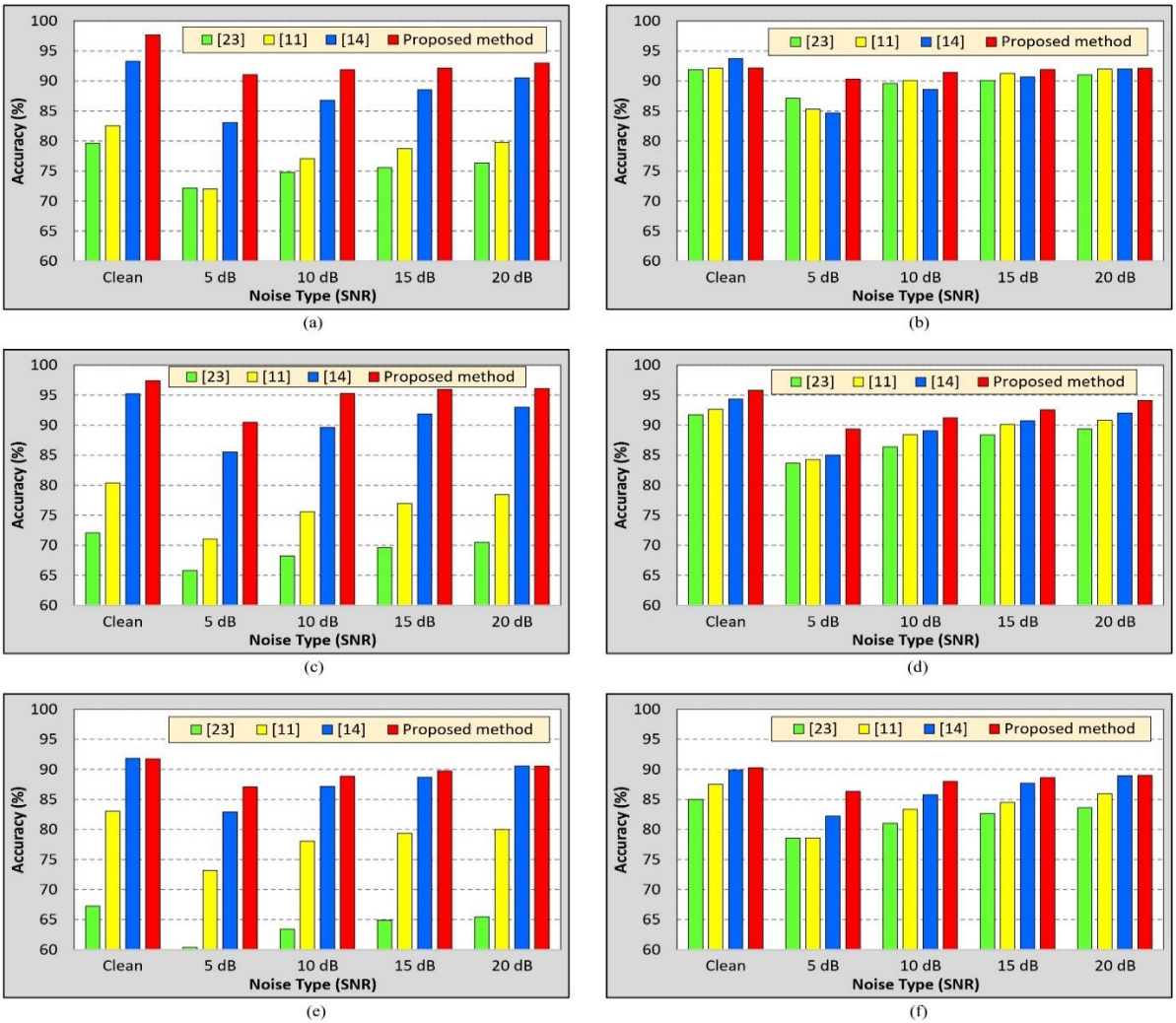
۴-۳- تنظیم پارامترها

در شبکه‌های عصبی پیچشی تمام پارامترهای مربوط به وزن‌ها به صورت تصادفی مقداردهی اولیه می‌شوند. برای آموزش گفتار از داده‌های تمیز و نویزی به صورت جداگانه استفاده شده است و در تمام حالت‌ها ساختار شبکه مانند ساختار پیشنهادی است. همچنین توابع فعالیت خروجی خطی بوده و از گرادیان نزولی تصادفی[۹] و از اندازه‌ی دسته‌های کوچک[۱۰] برای آموزش شبکه استفاده شده است. به دلیل وجود مقادیر مثبت و منفی در ورودی و خروجی، در لایه‌ی کانولوشن تابع فعالیت تانژانت هایپربولیک بکار گرفته شده است. تعداد گام‌های[۱۱] مورد نیاز برای آموزش شبکه 100 بوده و بعد از 100 تکرار دقت شبکه پایین می‌آید و در حقیقت overfitting رخ می‌دهد. اندازه دسته[۱۲] به 128 تنظیم شده و تعداد تکرار[۱۳] بر اساس اندازه داده های آموزشی تعین می شود. در نهایت نرخ یادگیری شبکه برای تمام حالات 0.0001 در نظر گرفته شده است. اینکه برای هر گوینده چند جمله به عنوان آموزش و چند جمله به عنوان تست در نظر گرفته شود جای بحث دارد. استانداردترین روش برای تقسیم‌بندی داده به آموزش و تست، روش ارزیابی k-fold نام دارد. در این روش داده‌ها به k دسته‌ی مساوی تقسیم می‌شوند و در k مرحله سیستم آموزش داده شده و تست می شود. در هر مرحله k-1 دسته برای آموزش و 1 دسته‌ی باقیمانده برای تست مورد استفاده قرار می‌گیرید. 10-flold از روش‌های مرجع برای ارزیابی است که در 10 مرحله سیستم آموزش داده شده و تست می‌شود که در اینجا از همین روش استفاده شده است. استخراج ویژگی در یک مرحله انجام می شود. سپس مدل CNN ۱۰ بار بر روی داده های آموز داده شده و تست می شود. نتایج هر حالت میانگین این ۱۰ اجرا بوده که در جداول گزارش می شوند.

۵- نتایج پیاده‌سازی

به منظور ارزیابی مدل پیشنهادی، نتایج پیاده‌سازی آن بر روی پایگاه‌های داده‌ی مذکور برای حالت‌های مختلف ارایه شده است. نتایج دقت روش پیشنهادی روی پایگاه داده‌ی Aurora2 در جدول ۷ و شکل (۱۳) آمده است.

[1] Subway
[2] Car
[3] Babble
[4] Exhibition
[5] Restaurant
[6] Street
[7] Airport
[8] Train Station
[9] Stochastic Gradient Descent (SGD)
[10] Mini Batch
[11] Epoch
[12] Batch Size
[13] Iteration



شکل ۱۳- نتایج دقت روی پایگاه داده‌ی Aurora2 برای حالت‌های مختلف، a) بخش A پایگاه داده و داده‌های Clean برای آموزش، b) بخش A پایگاه داده و داده‌های Noisy برای آموزش، c) بخش B پایگاه داده و داده‌های Clean برای آموزش، d) بخش B پایگاه داده و داده‌های Noisy برای آموزش، e) بخش C پایگاه داده و داده‌های Clean برای آموزش، f) بخش C پایگاه داده و داده‌های Noisy برای آموزش.

از نتایج جدول ۷ و شکل (۱۳) روی پایگاه داده‌ی Aurora2 داریم: ۱) روش پیشنهادی نسبت به روش‌های [۳۴]، [۱۲] و [۱۵]، از 30 آزمایش مختلف در 27 حالت (90.00%) به دقت بالاتری رسیده است؛ ۲) متوسط دقت روش پیشنهادی روی همه‌ی آزمایشات برابر 91.74% بوده است که نسبت به روش‌های [۳۴]، [۱۲] و [۱۵]، به ترتیب 2.61%، 8.98% و 13.54% بالاتر بوده است؛ ۳) بالاترین مقدار دقت متعلق به روش پیشنهادی بوده و برابر 97.66% است. نتایج دقت روش پیشنهادی روی پایگاه داده‌ی TIMIT برای حالت‌های سیگنال تمیز و نویزی در جدول ۷ و شکل (۱۴) آمده است. نویز با SNRهای 0 dB، 5 dB، 10 dB و 15 dB به سیگنال تمیز اضافه می‌شود که سیگنال نویز از پایگاه داده‌ی NOISEX-92 [۵۰] گرفته شده است. نتایج جدول ۸ و شکل (۱۴) روی پایگاه داده‌ی TIMIT را می‌توان به این صورت جمع‌بندی کرد: ۱) روش پیشنهادی نسبت به روش‌های [۳۴]، [۱۲] و [۱۵]، از 61 آزمایش مختلف در 48 حالت (78.69%) به دقت بالاتری رسیده است؛ ۲) عملکرد روش [۱۲] برای نویز Machingun بهتر از بقیه‌ی روش‌ها بوده است؛ ۳) متوسط دقت روش پیشنهادی روی همه‌ی آزمایشات برابر 93.97% بوده است که نسبت به روش‌های [۳۴]، [۱۲] و [۱۵]، به ترتیب 1.34%، 12.49% و 22.38% بالاتر بوده است. ۴) روش پیشنهادی در 10 حالت (66.67%) از 15 حالت مختلف نویز عملکرد بهتری نسبت به روش‌های [۳۴]، [۱۲] و [۱۵] داشته است.

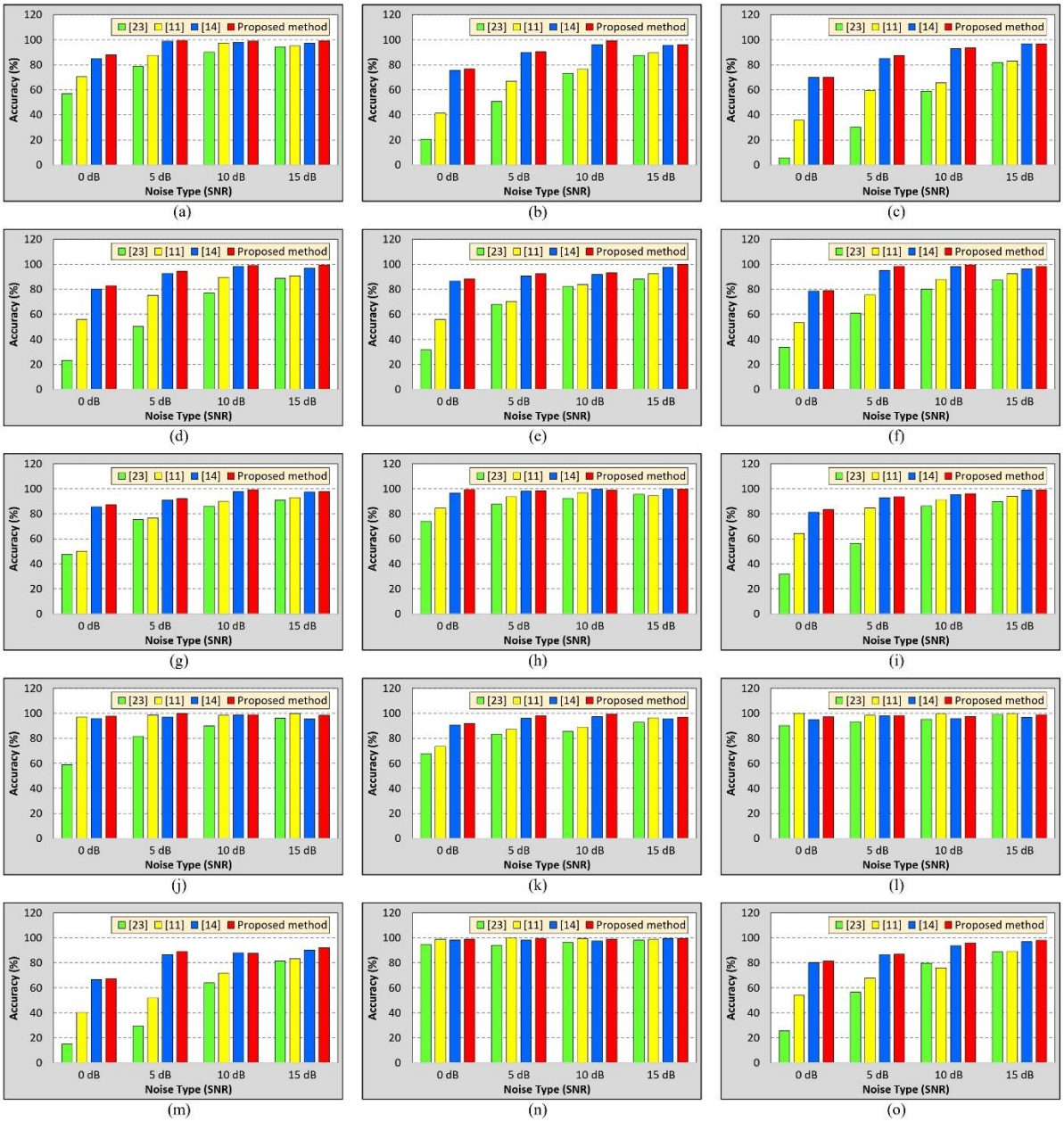
جدول ۷: نتایج دقت روی پایگاه داده‌ی Aurora2.

Dataset Section	Training Data	Noise Type (SNR)	[۳۴]	[۱۲]	[۱۵]	Proposed Method
A	Clean	Clean	79.61	82.52	93.28	**97.66**
		5 dB	72.15	72.00	83.09	**91.05**
		10 dB	74.79	77.07	86.81	**91.88**
		15 dB	75.55	78.73	88.53	**92.16**
		20 dB	76.34	79.76	90.49	**92.99**
A	Noisy	Clean	91.86	92.11	**93.71**	92.19
		5 dB	87.11	85.29	84.68	**90.32**
		10 dB	89.58	90.07	88.60	**91.44**
		15 dB	90.09	91.26	90.67	**91.94**
		20 dB	91.01	91.98	92.05	**92.09**
B	Clean	Clean	72.07	80.37	95.22	**97.42**
		5 dB	65.79	71.03	85.51	**90.48**
		10 dB	68.22	75.58	89.62	**95.29**
		15 dB	69.60	76.93	91.86	**95.94**
		20 dB	70.48	78.46	93.01	**96.12**
B	Noisy	Clean	91.75	92.64	94.31	**95.81**
		5 dB	83.66	84.27	85.04	**89.32**
		10 dB	86.41	88.41	89.05	**91.23**
		15 dB	88.38	90.14	90.73	**92.56**
		20 dB	89.39	90.82	92.02	**94.12**
C	Clean	Clean	67.21	83.04	**91.83**	91.76
		5 dB	60.35	73.17	82.91	**87.09**
		10 dB	63.40	78.03	87.18	**88.89**
		15 dB	64.85	79.36	88.67	**89.73**
		20 dB	65.44	79.96	**90.58**	90.57
C	Noisy	Clean	85.00	87.49	89.91	**90.26**
		5 dB	78.57	78.60	82.20	**86.31**
		10 dB	81.02	83.37	85.77	**88.00**
		15 dB	82.64	84.52	87.69	**88.63**
		20 dB	83.65	85.95	88.94	**89.01**
Average	-	-	78.20	82.76	89.13	**91.74**

جدول ۸: نتایج دقت روی پایگاه داده‌ی TIMIT.

Noise Type	Noise Level	[۳۴]	[۱۲]	[۱۵]	Proposed Method
Clean	-	97.10	97.64	97.25	**98.84**
Babble	0 dB	56.93	70.45	85.03	**87.92**
	5 dB	78.76	87.65	98.88	**99.42**
	10 dB	90.13	97.39	97.93	**99.01**
	15 dB	94.22	95.10	97.20	**99.29**
Buccaneer1	0 dB	20.51	41.33	75.70	**76.68**
	5 dB	50.83	66.99	90.06	**90.55**
	10 dB	73.23	76.50	96.23	**99.21**
	15 dB	87.52	89.61	95.67	**95.95**
Buccaneer2	0 dB	5.68	35.67	69.91	**70.07**
	5 dB	30.16	59.18	85.16	**87.31**
	10 dB	58.93	65.55	92.78	**93.63**
	15 dB	81.92	82.98	**96.91**	96.78
Destroyer Engine	0 dB	23.25	55.84	80.32	**82.82**
	5 dB	50.29	75.12	92.81	**94.50**
	10 dB	77.12	89.56	98.27	**99.05**
	15 dB	88.75	90.76	97.07	**99.63**
Destroyer Ops	0 dB	31.52	55.91	86.63	**88.33**
	5 dB	67.99	70.33	90.63	**92.40**
	10 dB	82.24	83.81	91.81	**93.32**
	15 dB	88.09	92.46	97.61	**99.86**
F16	0 dB	33.70	53.25	78.59	**79.03**
	5 dB	61.03	75.49	95.11	**98.43**
	10 dB	80.06	87.84	98.16	**99.61**
	15 dB	87.30	92.54	96.38	**98.41**
Factory1	0 dB	47.50	50.13	85.60	**87.35**
	5 dB	75.30	76.64	90.95	**92.16**
	10 dB	86.15	89.89	97.70	**99.27**
	15 dB	90.78	93.00	97.46	**97.88**
Factory2	0 dB	73.84	84.55	96.73	**99.32**
	5 dB	87.70	93.74	98.14	**98.34**
	10 dB	92.38	96.78	**99.54**	98.81
	15 dB	95.56	94.42	**99.88**	99.62
Hfchannel	0 dB	31.81	64.32	81.30	**83.39**
	5 dB	56.33	84.66	92.85	**93.75**
	10 dB	86.27	91.23	95.43	**96.09**
	15 dB	89.70	94.16	99.01	**99.12**
Leopard	0 dB	58.91	97.00	95.82	**97.69**
	5 dB	81.51	98.55	97.02	**99.81**
	10 dB	89.86	98.33	98.64	**98.88**
	15 dB	95.99	**99.88**	95.55	98.35
M109	0 dB	67.73	73.49	90.76	**91.81**
	5 dB	83.33	87.17	95.90	**98.20**
	10 dB	85.45	88.91	97.27	**99.23**
	15 dB	92.81	96.34	95.75	**96.88**
Machinegun	0 dB	90.20	**99.98**	95.09	97.17
	5 dB	93.16	**98.35**	97.87	98.29
	10 dB	95.13	**99.51**	95.81	97.43
	15 dB	98.95	**99.57**	96.74	98.87
Volvo	0 dB	14.96	40.13	66.64	**67.42**
	5 dB	29.49	52.04	86.49	**89.20**
	10 dB	63.96	71.76	**87.93**	87.55
	15 dB	81.52	83.23	90.19	**92.09**
Pink	0 dB	94.70	**98.78**	98.21	98.76
	5 dB	93.99	**99.82**	98.08	99.29
	10 dB	96.20	**99.37**	97.46	98.76
	15 dB	98.09	98.64	**99.50**	99.46
White	0 dB	25.67	53.95	80.15	**81.40**
	5 dB	56.52	67.74	86.64	**86.88**
	10 dB	79.73	75.98	93.57	**95.96**
	15 dB	88.72	89.16	96.83	**97.88**
Average	-	71.59	81.48	92.63	**93.97**

با مقایسه‌ی نتایج به دست آمده روی دو پایگاه داده مذکور، مشاهده می‌شود که: ۱) روش پیشنهادی روی پایگاه‌های داده Aurora2 و TIMIT به ترتیب در 90.00% و 78.69% از حالت‌های مختلف، به دقت بالاتری نسبت به سایر روش‌ها رسیده است و این نشان می‌دهد که روش پیشنهادی نسبت به سایر روش‌ها، روی پایگاه داده Aurora2 عملکرد بهتری داشته است. ۲) برای سیگنال clean، دقت روش پیشنهادی روی پایگاه داده Aurora2 برابر 97.66% و روی پایگاه داده TIMIT برابر 98.84% بوده است بنابراین روش پیشنهادی روی پایگاه داده TIMIT به دقت بالاتری دست یافته است. ۳) در نویزهای 5dB، 10dB و 15dB، بالاترین دقت پایگاه داده Aurora2 به ترتیب برابر 91.05%، 95.29% و 95.94% و بالاترین دقت پایگاه داده TIMIT به ترتیب برابر 99.42%، 99.61% و 99.86% است بنابراین برای نویز 5dB، 10dB



شکل ۱۴- نتایج دقت روی پایگاه داده‌ی TIMIT برای نویزهای مختلف؛ a) Babble، b) Buccaneer1، c) Buccaneer2، d) Destroyer Engine، e) Destroyer Ops، f) F16، g) Factory 1، h) Factory 2، i) Hfchannel، j) Leopard، k) M109، l) Machinegun، m) Volvo، n) Pink، o) White.

و 15dB نیز دقت روش پیشنهادی روی پایگاه داده TIMIT بالاتر از Aurora2 بوده است.

۶-نتیجه‌گیری

در این مقاله، یک سیستم تشخیص گوینده مبتنی بر ویژگی‌های زمان-فرکانس گابور و شبکه‌های عصبی کانولوشن پیشنهاد شد. در روش پیشنهادی، ابتدا اسپکتروگرام سیگنال تشکیل شد، سپس با طراحی فیلترهای گابور و عبور سیگنال از این فیلترها، ویژگی‌های سیگنال استخراج شد. در نهایت با استفاده از شبکه‌ی عصبی کانولوشن، امکان شناسایی گوینده فراهم شد. پیاده سازی و ارزیابی روش پیشنهادی بر روی دو پایگاه داده‌ی Aurora2 و TIMIT نشان داد که دقت روش پیشنهادی قابل رقابت با روش‌های پیشین است.

مراجع

[1] Dunn, J. S., and Podio, F. Biometrics Consortium website, http:// www.biometrics.org.

[2] Campbell, Joseph P. “Speaker recognition: A tutorial”, Proceedings of the IEEE, Vol. 85, No. 9, 1997, pp. 1437–1462.

[3] Müller, C., Speaker Classification I: Fundamentals, Features, and Methods, Springer-Verlag Berlin Heidelberg, 2007.

[4] Müller, C., Speaker Classification II, Springer-Verlag Berlin Heidelberg, 2007.

[5] Keshet, J., and Bengio, S. (Eds.), Automatic speech and speaker recognition: large margin and kernel methods, John Wiley & Sons, 2009.

[6] Ohi, A. Q., Mridha, M. F., Hamid, M. A., and Monowar, M. M., “Deep speaker recognition: process, progress, and challenges”, IEEE Access, Vol. 9, 2021, pp. 89619-89643.

[7] Hanifa, R. M., Isa, K., and Mohamad, S., “A review on speaker recognition: Technology and challenges”, Computers & Electrical Engineering, Vol. 90, 2021, pp. 107005.

[8] Lan, J., Zhang, R., Yan, Z., Wang, J., Chen, Y., and Hou, R., “Adversarial attacks and defenses in Speaker Recognition Systems: A survey”, Journal of Systems Architecture, Vol. 127, 2022, pp. 102526.

[9] Schädler, M. R., Meyer, B. T., and Kollmeier, B., “Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition”, The Journal of the Acoustical Society of America, Vol. 131, No. 5, 2012, pp. 4134-4151.

[10] Mesgarani, N., Slaney, M., and Shamma, S. A., “Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations”, IEEE Transactions on Audio, Speech, and Language Processing, Vol. 14, No. 3, 2006, pp. 920-930.

[11] Sahidullah, M., and Saha, G., “A novel windowing technique for efficient computation of MFCC for speaker recognition”, IEEE signal processing letters, Vol. 20, No. 2, 2012, pp. 149-152.

[12] Qi, M., Yu, Y., Tang, Y., Deng, Q., Mai, F., and Zhaxi, N., “Deep CNN with se block for speaker recognition”, In 2020 Information Communication Technologies Conference (ICTC), IEEE, 2020, pp. 240-244.

[13] Ghalamiosgouei, S., and Geravanchizadeh, M., “Robust Speaker Identification Based on Binaural Masks”, Speech Communication, Vol. 132, 2021, pp. 1-9.

[14] Chakroun, R., and Frikha, M., “Robust text-independent speaker recognition with short utterances using Gaussian mixture models”, In 2020 International Wireless Communications and Mobile Computing (IWCMC), IEEE, 2020, pp. 2204-2209.

[15] Moumin, A. A., and Kumar, S. S., “Automatic Speaker Recognition using Deep Neural Network Classifiers”, In 2021 2nd International Conference on Computation, Automation and Knowledge Management (ICCAKM), IEEE, 2021, pp. 282-286.

[16] Lin, T., and Zhang, Y., “Speaker recognition based on long-term acoustic features with analysis sparse representation”, IEEE Access, Vol. 7, 2019, pp. 87439-87447.

[17] Jiahong, L., Jie, B., Yingshuang, C., and Chun, L., “An Adaptive ResNet Based Speaker Recognition in Radio Communication”, In 2021 IEEE International Conference on Emergency Science and Information Technology (ICESIT), 2021, pp. 161-164.

[18] Prachi, N. N., Nahiyan, F. M., Habibullah, M., and Khan, R., “Deep Learning Based Speaker Recognition System with CNN and LSTM Techniques”, In 2022 Interdisciplinary Research in Technology and Management (IRTM), 2022, pp. 1-6.

[19] Wang, Y., Wan, S., Zhang, S., and Yu, J., “Speaker recognition of fiber-optic external Fabry-Perot interferometric microphone based on Deep Learning”, IEEE Sensors Journal, Vol. 22, No. 13, 2022, pp. 12906-12912.

[20] Balpande, M., Sansare, R., Padelkar, T., and Shinde, V., “Speaker Recognition based on Mel-Frequency Cepstral Coefficients and Vector Quantization”, In 2021 IEEE Bombay Section Signature Conference (IBSSC), 2021, pp. 1-6.

[21] Roy, M. K., and Keshwala, U., "Res2Net based Text Independent Speaker recognition system", In 2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2022, pp. 612-616.

[22] Wang, R., Ao, J., Zhou, L., Liu, S., Wei, Z., Ko, T., ... and Zhang, Y., "Multi-View Self-Attention Based Transformer for Speaker Recognition", In 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 6732-6736.

[23] Orken, M., Dina, O., Keylan, A., Tolganay, T., and Mohamed, O., "A study of transformer-based end-to-end speech recognition system for Kazakh language", Scientific Reports, Vol. 12, No. 1, 2022, pp. 1-11.

[24] Faúndez-Zanuy, M., "Speaker recognition by means of a combination of linear and nonlinear predictive models", arXiv preprint arXiv:2203.03190.

[25] Hu, H. R., Song, Y., Liu, Y., Dai, L. R., McLoughlin, I., and Liu, L., "Domain Robust Deep Embedding Learning for Speaker Recognition", In 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 7182-7186.

[26] Chowdhury, A., Cozzo, A., and Ross, A., "Domain Adaptation for Speaker Recognition in Singing and Spoken Voice", In 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 7192-7196.

[27] Bahmaninezhad, F., Zhang, C., and Hansen, J. H., "An investigation of domain adaptation in speaker embedding space for speaker recognition", Speech Communication, Vol. 129, 2021, pp. 7-16.

[28] Bharath, K. P., and Kumar, R., "Multitaper based MFCC feature extraction for robust speaker recognition system", In 2019 Innovations in Power and Advanced Computing Technologies (i-PACT), IEEE, Vol. 1, 2019, pp. 1-5.

[29] Nunes, J. A. C., Macêdo, D., and Zanchettin, C., "Am-mobilenet1d: A portable model for speaker recognition", In 2020 International Joint Conference on Neural Networks (IJCNN), IEEE, 2020, pp. 1-8.

[30] Nunes, J. A. C., Macêdo, D., and Zanchettin, C., "Additive margin sincnet for speaker recognition", In 2019 International Joint Conference on Neural Networks (IJCNN), IEEE, 2019, pp. 1-5.

[31] Liu, Z., Wu, Z., Li, T., Li, J., and Shen, C., "GMM and CNN hybrid method for short utterance speaker recognition", IEEE Transactions on Industrial informatics, Vol. 14, No. 7, 2018, pp. 3244-3252.

[32] Dai, M., Dai, G., Wu, Y., Xia, Y., Shen, F., and Zhang, H., "An Improved Feature Fusion for Speaker Recognition", In 2019 IEEE Fourth International Conference on Data Science in Cyberspace (DSC), IEEE, 2019, pp. 183-187.

[33] Avila, A. R., O'Shaughnessy, D., and Falk, T. H., "Automatic speaker verification from affective speech using Gaussian mixture model based estimation of neutral speech characteristics", Speech Communication, Vol. 132, 2021, pp. 21-31.

[34] Rashno, E., Akbari, A., and Nasersharif, B., "A convolutional neural network model based on neutrosophy for noisy speech recognition", In 2019 4th International Conference on Pattern Recognition and Image Analysis (IPRIA), IEEE, 2019, pp. 87-92.

[35] Bian, T., Chen, F., and Xu, L., "Self-attention based speaker recognition using Cluster-Range Loss", Neurocomputing, 2019, pp. 368, 59-68.

[36] Devi, K. J., Singh, N. H., and Thongam, K., "Automatic speaker recognition from speech signals using self organizing feature map and hybrid neural network", Microprocessors and Microsystems, Vol. 79, 2020, pp.103264.

[37] Chien, J. T., and Peng, K. T., "Neural adversarial learning for speaker recognition", Computer Speech & Language, Vol. 58, 2019, pp. 422-440.

[38] Han, J. H., Bae, K. M., Hong, S. K., Park, H., Kwak, J. H., Wang, H. S., ... and Lee, K. J., "Machine learning-based self-powered acoustic sensor for speaker recognition", Nano Energy, Vol. 53, 2018, pp. 658-665.

[39] Zhang, X., Zou, X., Sun, M., Zheng, T. F., Jia, C., and Wang, Y., "Noise robust speaker recognition based on adaptive frame weighting in GMM for i-vector extraction", IEEE Access, Vol. 7, 2019, pp. 27874-27882.

[40] Chowdhury, A., and Ross, A., "Fusing MFCC and LPC features using 1D triplet CNN for speaker recognition in severely degraded audio signals", IEEE transactions on information forensics and security, Vol. 15, 2019, pp. 1616-1629.

[41] Xu, J., Li, S., Jiang, J., and Dou, Y., “A simplified speaker recognition system based on FPGA platform”, IEEE Access, Vol. 8, 2019, pp. 1507-1516.

[42] Mesgarani, N., David, S. V., Fritz, J. B., and Shamma, S. A., “Phoneme representation and classification in primary auditory cortex”, The Journal of the Acoustical Society of America, Vol. 123, No. 2, 2008, pp. 899-909.

[43] Ezzat, T., Bouvrie, J. V., and Poggio, T. A., “Spectro-temporal analysis of speech using 2-d Gabor filters”, In Interspeech, 2007, pp. 506-509.

[۴۴] سیاوش حسینی، سعید ستایشی، غلامحسین روشنی، عبدالحمید زاهدی و فرزین شماع، "افزایش کارآیی جریان سنج دوفازی با استفاده از روش های استخراج ویژگی حوزه ی فرکانس و شبکه عصبی در طیف خروجی آشکارساز"، مدل سازی در مهندسی، دوره ۱۹، شماره ۶۷، زمستان ۱۴۰۰، صفحه ۴۷-۵۷.

[۴۵] میثم عفتی، رحمت مدندوست، و زینب فلاح زرجو بازکیایی، "ارزیابی عملکرد مدل های شبکه عصبی مصنوعی، نروفازی و رگرسیون چند متغیره در پیش بینی مقاومت فشاری بتن به کمک روش بارنقطه ای"، مدل سازی در مهندسی، دوره ۱۸، شماره ۶۲، پاییز ۱۳۹۹، صفحه ۹۹-۱۱۳.

[۴۶] محمدجسین ولایتی، "ارزیابی قابلیت ضریب مشارکت ژنراتورها به منظور تعیین نوع نوسانات سیگنال کوچک سیستم قدرت با استفاده از روش‌های تحلیلی و پیش‌بینی همزمان آن‌ها با استفاده از شبکه عصبی"، مدل سازی در مهندسی، دوره ۱۳، شماره ۴۲، پاییز ۱۳۹۴، صفحه ۱۱۹-۱۳۳.

[47] TIMIT dataset, available online on: https://catalog.ldc.upenn.edu/LDC93S1. Last accessed at 14 September 2021.

[48] The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions.

[49] Naing, H. M. S., Hidayat, R., Hartanto, R., and Miyanaga, “Discrete wavelet denoising into MFCC for noise suppressive in automatic speech recognition system”, International Journal of Intelligent Engineering and Systems, Vol. 13, No. 2, 2020, pp. 74-82.

[50] NOISEX-92 noise dataset, available online on: http://spib.linse.ufsc.br/noise.html. Last accessed at 14 September 2021.