مقاله مروری

مروری بر استفاده از فیلترهای پارامتری در معماری‌های ژرف برای کاربردهای پردازش گفتار

حسین فیاضی
دانشجوی دکترا
آزمایشگاه پردازش هوشمند صدا، دانشکده مهندسی و علوم کامپیوتر، دانشگاه شهید بهشتی
h_fayyazi@sbu.ac.ir

یاسر شکفته*
استادیار
آزمایشگاه پردازش هوشمند صدا، دانشکده مهندسی و علوم کامپیوتر، دانشگاه شهید بهشتی
y_shekofteh@sbu.ac.ir

تاریخ دریافت: ۱۴۰۳/۲/۲۹ تاریخ پذیرش: ۱۴۰۳/۶/۵

چکیده

در روش‌های پردازش گفتار سنتی، عملیات استخراج ویژگی و دسته‌بندی در دو مرحله جداگانه انجام می‌شد. با گسترش استفاده از شبکه‌های عصبی ژرف، روش‌هایی ارائه شدند که در آن مدل‌سازی ارتباط بین مشخصه‌های آکوستیک و آوایی گفتار و دسته‌بندی آن به‌صورت همزمان از روی سیگنال زمانی گفتار انجام می‌شد. لایه اول پیچشی این شبکه‌ها را می توان به‌عنوان یک بانک فیلتر در نظر گرفت که هر کدام از فیلترها نسبت به باندهای فرکانسی متفاوتی حساس هستند. پس‌ازآن، با هدف افزودن قابلیت تفسیرپذیری و کاهش تعداد پارامترهای شبکه، استفاده از فیلترهای پارامتری مورد توجه قرار گرفت. معماری سینک نت[1] که در سال ۲۰۱۸ برای کاربرد شناسایی گوینده و شناسایی گفتار ارائه شد، مهم‌ترین تلاش در این زمینه بود. در لایه اول پیچشی این معماری، به‌جای فیلترهایی که تمام وزن‌های آن قابل یادگیری هستند، فیلترهای میان‌گذر مستطیلی یاد گرفته می‌شد. ازآنجاکه این فیلترها با تعداد کمی پارامتر قابل مدل‌سازی بودند، افزودن این محدودیت به شبکه باعث شد تعداد پارامترهای شبکه کمتر شده و سرعت همگرایی و دقت آن افزایش یابد. بعلاوه، با تجزیه‌وتحلیل بانک فیلتری که توسط مدل شبکه عصبی یاد گرفته می‌شد، اطلاعات ارزشمندی از نحوه عملکرد مدل به دست می‌آمد. کاهش تعداد پارامترهای شبکه و افزایش دقت و قدرت تفسیرپذیری مدل باعث شده است که امروزه استفاده از انواع دیگر فیلترهای پارامتری در کاربردهای مختلف پردازش گفتار مورد توجه قرار گیرد. در این مقاله انواع فیلترهای پارامتری معرفی شده و نحوه استفاده از آن‌ها در معماری‌های عمیق مختلف شرح داده می‌شوند. در انتها نیز کاربردهایی از پردازش گفتار که در آن‌ها از این فیلترها استفاده شده است، معرفی شده‌اند.

واژگان کلیدی: پردازش گفتار، یادگیری ژرف، تفسیرپذیری، فیلترهای پارامتری، سینک نت

۱. مقدمه

گفتار به‌عنوان ابزار اولیه ارتباط بین انسان‌ها، روشی برای بیان نیازها، انتقال مفاهیم ذهنی و احساسات درونی است. پردازش گفتار زمینه تحقیقاتی است که در آن از روش‌های پردازش سیگنال و یادگیری ماشین برای بهبود ارتباط انسان و ماشین از طریق گفتار استفاده می‌شود. پردازش گفتار، شامل زیرشاخه‌های کاربردی زیادی مانند بازشناسی گفتار، تبدیل متن به گفتار، بازشناسی گوینده، تبدیل صدا، بهسازی گفتار، تعیین جنسیت گوینده، تشخیص احساس و تشخیص ناهنجاری‌های گفتاری است. با توسعه هوش مصنوعی و استفاده از روش‌های مبتنی بر یادگیری ژرف، عملکرد سامانه‌های پردازش گفتار بهبود قابل توجهی پیدا کرده‌اند. دراین‌بین، استفاده از معماری های عمیق مانند شبکه عصبی پیچشی[۲]، مدل‌های مبتنی بر سازوکار توجه[۳]، مدل‌های مبتنی بر شبکه‌های عصبی بازگشتی[۴]، مدل‌های مبتنی بر معماری‌های رقابتی مولد[۵] و اخیراً نیز معماری‌های مبتنی بر مبدل‌ها[۶] مورد توجه قرار گرفته‌اند.

در بسیاری از کاربردهای پردازش گفتار، ویژگی‌هایی مانند طیف نما به‌عنوان ورودی یک شبکه عصبی پیچشی مورد استفاده قرار می‌گیرند. اگرچه طیف نما نسبت به ویژگی‌های دستی، اطلاعات بیشتری دارد اما برای رسیدن به دقت مناسب بایستی ابرپارامترهای مختلفی همچون طول قاب، میزان هم‌پوشانی، شکل پنجره و تعداد بین‌های فرکانسی تنظیم شوند. این مساله باعث شد که محققان به‌صورت مستقیم از سیگنال زمانی گفتار یا اصطلاحاً سیگنال خام به‌عنوان ورودی استفاده و فرایند استخراج ویژگی را به طور کامل به مدل‌های ژرف واگذار کنند [۷].

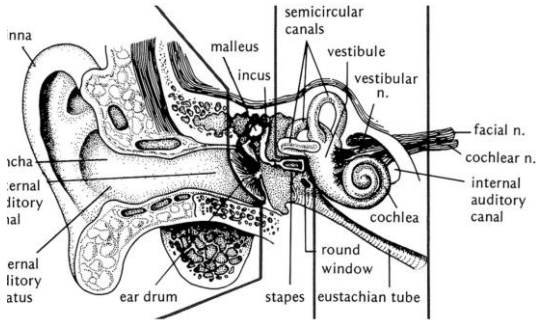
یکی از مهم‌ترین محدودیت‌های روش‌های یادگیری ژرف، آن است که این روش‌ها ماهیتی جعبه-سیاه[۷] دارند؛ به‌بیان‌دیگر در مورد چگونگی رسیدن به تصمیم نهایی توضیحی داده نمی‌شود. این امر نه‌تنها اعتبار نتیجه تصمیم شبکه را به چالش می‌کشد بلکه می‌تواند مانعی برای بهبود معماری‌های ژرف و بسط آن به کاربردهای دیگر باشد.

اخیراً، زمینه تحقیقاتی هوش مصنوعی تفسیرپذیر[۸] (XAI)، برای حل این مشکل مورد توجه قرار گرفته است. تفسیرپذیری[۹] توانایی توضیح یا ارائه رفتار مدل به‌صورت قابل درک برای انسان است [۸]. XAI می‌تواند به شناسایی نقاط قوت و ضعف مدل کمک کند. همچنین می‌تواند به روشن شدن مسیر تحقیقاتی آینده کمک کرده و باعث توسعه معماری‌های ژرف با قابلیت اطمینان بیشتر شود [۹]. در تحقیقات مختلف برای درک رفتار مدل، دو رهیافت کلی پیشنهاد شده است. رهیافت اول، استفاده از مدل‌هایی است که خود ماهیتی تفسیرپذیر دارند. مدل‌های خطی و درخت تصمیم نمونه‌هایی از این مدل‌ها هستند. اما این مدل‌ها معمولاً کارایی پایینی دارند. بعلاوه، همواره در بکارگیری مدل‌ها، مصالحه‌ای بین دقت و تفسیرپذیری وجود دارد. مدل‌های ساده‌تر تفسیرپذیری بالاتری دارند اما دقت آن‌ها پایین است و هر چه مدل پیچیده‌تر می‌شود، دقت آن بالاتر و تفسیرپذیری آن پایین می‌آید. در مواردی که برای داشتن کارایی بالا ناگزیر به استفاده از مدل‌های پیچیده‌تر هستیم، از رهیافت دوم که توضیح مدل به روش آزمون تقیبی[۱۰] است، استفاده می‌شود. در این رهیافت، توضیحاتی در مورد رفتار مدل ارائه می‌شود که دید توصیفی مناسبی از عملکرد داخلی آن در اختیار قرار می‌دهد [۱۰].

در حوزه کاربردی پردازش گفتار تلاش‌های ارزشمندی برای کمک به تفسیرپذیری مدل‌های ژرف شده است. یادگیری بازنمایی تفسیرپذیر با استفاده از روش وزن‌دهی رابطه [۱۱]، استفاده از تبدیل خودرمزگذار[11] [۱۲]، استفاده از SHAP[12] برای توضیح رفتار شبکه‌های ژرف در کاربرد تشخیص جعل ژرف صوتی[13] [۹]، و کمک به شبکه‌های عصبی پیچشی برای یافتن فیلترهای معنی‌دار در [۷]، نمونه‌هایی از این تحقیقات هستند.

نتایج ارائه شده در مرجع [۷] نشان می‌دهند که لایه اول پیچشی یکی از کلیدی‌ترین بخش‌های شبکه‌هایی است که از سیگنال خام به‌عنوان ورودی استفاده می‌کنند. ورودی این لایه نه‌تنها داده‌هایی با ابعاد بسیار بالا است بلکه مستعد مواجه شدن با مساله محوشدگی گرادیان، به‌خصوص در معماری‌های ژرف‌تر، است. علاوه بر این، فیلترهایی که در این لایه یاد گرفته می‌شوند، شکلی نویزی و چندباندی[14] خواهند داشت؛ این مساله در زمانی که تعداد داده‌های آموزشی کم باشد، بیشتر خود را نشان می‌دهد. بنابراین بهتر آن است که شبکه‌ای طراحی شود که ویژگی‌ها را بر اساس مشخصه‌های سیگنال ورودی استخراج کند. یکی از روش‌هایی که برای حل این مساله ارائه شده است، استفاده از فیلترهایی با شکل مشخص در لایه اول معماری‌های ژرف است.

در بخش دوم این مقاله، مدل فیلتری که در گوش انسان اتفاق می‌افتد، توضیح داده می‌شود. در بخش سوم با انواع فیلترهایی که در معماری‌های ژرف مختلف برای کاربردهای پردازش گفتار مورد استفاده قرار گرفته‌اند، آشنا می‌شویم. پس از آن در بخش چهارم، معماری‌هایی که از این فیلترها استفاده کرده‌اند، معرفی



شکل ۱. ساختار سیستم شنیداری انسان [۱]

شده و در بخش پنجم مهم‌ترین کاربردهایی از پردازش گفتار که از این فیلترها و معماری‌ها استفاده کرده‌اند، به اختصار ارائه می‌شوند. در انتها نیز خلاصه و نتیجه‌گیری مقاله ارائه می‌شود.

۲. مدل فیلتری شنیداری

بیشتر پیشرفت‌هایی که در زمینه هوش مصنوعی صورت گرفته است، مرهون تقلید از عملکرد بخش‌های مختلف جسمی یا رفتاری موجودات زنده است. در این بخش، در ابتدا مکانیزم شنیداری انسان بر اساس توضیحاتی که در [۱۳] آمده است، توضیح داده می‌شود. سپس نشان خواهیم داد، این سازوکار چگونه می‌تواند به ساختار شبکه‌های عصبی پیچشی نگاشت شود.

قسمت جانبی سیستم شنیداری بیشتر پستانداران مشابه یکدیگر است. شکل ۱ ساختار سیستم شنیداری انسان را نمایش می‌دهد. این قسمت شامل گوش خارجی، گوش میانی و گوش داخلی است. گوش خارجی از دو بخش لاله گوش و کانال شنیداری تشکیل شده‌است. لاله گوش، صداهای ورودی، به‌خصوص فرکانس‌های بالا را به‌طور قابل‌ملاحظه‌ای تغییر می‌دهد. امواج صوتی از طریق کانال صوتی به پرده گوش رسیده و باعث ارتعاش آن می‌شوند. این

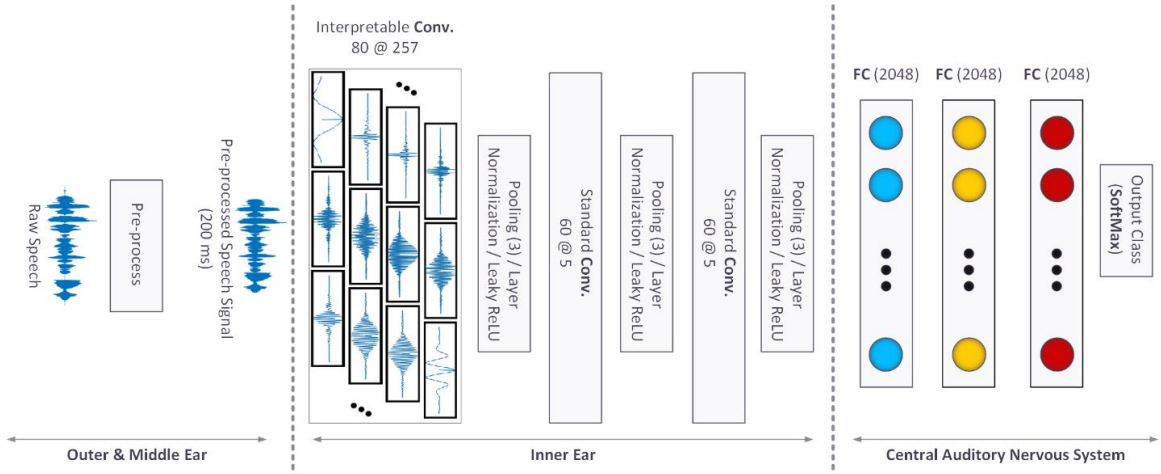
ارتعاش‌ها به‌وسیله سه استخوان کوچک (استخوانچه) از طریق گوش میانی به یک حفره پوشیده از غشا در دیواره استخوانی گوش داخلی می‌رسد. این حفره که پنجره بیضوی نام دارد، درواقع بخشی از حلزونی گوش است که ساختاری مارپیچی شکل داشته و درون آن از مایع غیرقابل فشرده‌سازی پر شده است.

درک عملکرد حلزونی جنبه‌های مختلف ادراک شنیداری انسان را روشن‌تر می‌کند. حلزونی گوش با دو غشا به نام‌های رایسنر[15] و بازیلار[16] پوشیده شده است. وقتی پنجره بیضوی به واسطه حرکت استخوانچه حرکت می‌کند، اختلاف فشاری در سراسر غشای بازیلار (BM) ایجاد می‌شود که باعث حرکت آن می‌شود. این اختلاف فشار و الگوی حرکت روی BM، مدتی طول می‌کشد تا منتشر شود و در طول BM متفاوت است. پاسخ BM به صداهای با فرکانس‌های مختلف، متأثّر از مشخصه‌های مکانیکی صدا است که در قسمت ابتدا و انتهای حلزونی متفاوت است. در قسمت ابتدایی، پاسخ BM باریک و تیز است و در قسمت انتهایی پاسخ آن عریض‌تر می‌شود. در نتیجه، مکان قله در الگوی ارتعاش‌ها، متناسب با فرکانس تحریک متفاوت است. در حقیقت، حلزونی گوش به مانند یک تجزیه‌کننده فوریه با توان آنالیز فرکانسی محدودتر، عمل می‌کند. فرکانسی که بیشترین پاسخ را در یک نقطه روی BM ایجاد می‌کند، فرکانس مشخصه[17] (CF) برای آن مکان نامیده می‌شود. در پاسخ به یک سیگنال سینوسی تک فرکانس، هر نقطه از BM تقریباً یک حرکت سینوسی با همان فرکانس موج ورودی خواهد داشت. اما لرزش برخی از قسمت‌های BM دارای دامنه بزرگ‌تری بوده و بعلاوه فاز لرزش در نقاط مختلف BM متفاوت است. بر این اساس، هر نقطه روی BM را می‌توان یک فیلتر میان‌گذر در نظر گرفت که فرکانس مرکزی و پهنای باند متفاوتی دارند. پاسخ BM در قسمت‌های مختلف دارای پهنای باند یکسانی نیست؛ به‌طوری‌که هر چه CF بیشتر می‌شود، پهنای باند نیز بزرگ‌تر می‌شود.

براساس مطالبی که تا این قسمت بیان شد، می‌توان گفت که سیستم شنیداری انسان قابلیت آنالیز فرکانسی دارد. به‌عبارت‌دیگر می‌تواند مولفه‌های سینوسی سیگنال صدا را از یکدیگر تفکیک کند. تفکیک‌پذیری یا آنالیز فرکانسی که به آن انتخاب فرکانسی[18] هم گفته می‌شود، معمولاً با مفهومی به نام ماسک کردن[19] بررسی می‌شود. این که یک تُن (صدای تک فرکانس) یا صدا به‌واسطه بلند بودن شدت صوتی تُن/صدای دیگر، غیرقابل شنیدن شود را ماسک کردن گویند که به علت افزایش حد آستانه قابلیت شنیده شدن آن تُن/صدا است.

از این‌رو یک سیگنال می‌تواند به‌راحتی با صدایی که مؤلفه‌های فرکانسی نزدیک به آن دارد، ماسک شود. بنابراین توانایی انسان در جداسازی مولفه‌های یک سیگنال، تا حدی به آنالیز فرکانسی که در BM رخ می‌دهد، وابسته است. بعلاوه می‌توان با استفاده از ماسک کردن، محدوده‌های آنالیز فرکانسی انجام شده در حلزونی را مشخص کرد. اگر تفکیک‌پذیری فرکانسی گوش در حدی نباشد که بتواند سیگنال را از ماسک‌کننده جدا کند، آنگاه ماسک کردن رخ داده است. بنابراین می‌توان از ماسک کردن برای کمی سازی آنالیز فرکانسی انجام شده در گوش استفاده کرد.

در یک آزمایش کلاسیک که در سال ۱۹۴۰ توسط فلتچر[20] انجام شده است، حد آستانه تشخیص سیگنال سینوسی به‌عنوان تابعی از پهنای باند یک ماسک‌کننده، اندازه‌گیری شد. بر اساس نتایج این آزمایش، فلتچر با



شکل ۲. معماری مدل سینک‌نت و نگاشت آن به سیستم شنیداری انسان [۶]

پیروی از هلمهولتز[۲۱]، این پیشنهاد را داد که سیستم شنیداری انسان طوری عمل می‌کند که شامل بانکی از فیلترهای میان‌گذر است که با یکدیگر هم‌پوشانی دارند. به این فیلترها، فیلترهای شنیداری[۲۲] گویند. هر مکان در BM به یک بازه محدود از فرکانس‌ها پاسخ می‌دهد، بنابراین هر نقطه متناظر با فیلتری با فرکانس مرکزی متفاوتی است.

معماری کلی یک شبکه پیچشی ژرف را می‌توان با سیستم شنوایی انسان تطبیق داد. اولین بخش از معماری مدل که شامل مراحل پیش‌پردازش است، منطبق با مکانیزم فیلتری گوش خارجی و میانی در سیستم شنیداری انسان است. تجزیه و تحلیل طیفی که در حلزونی گوش داخلی انجام می‌شود، با لایه‌های پیچشی این معماری منطبق است. در نهایت، لایه‌های تمام متصل، منطبق با سیستم عصبی شنیداری مرکزی مغز است. شکل ۲، این تطابق را برای شبکه سینک‌نت نمایش می‌دهد [۷].

۳. معرفی و مقایسه فیلترهای پارامتری

همان‌طور که در مقدمه بیان شد، زمانی که از سیگنال خام بعنوان ورودی یک شبکه عصبی پیچشی استفاده می‌شود، اولین لایه پیچشی مانند یک بانک فیلتر از فیلترهای خطی نامتغیر با زمان[۲۳] (LTI) عمل می‌کند که پاسخ ضربه آنها از روی داده‌ها یاد گرفته شده است [۱۴]. در این حالت چنانچه $x[n]$ سیگنال یک‌بعدی ورودی و $h[n]$ پاسخ ضربه فیلتر باشد، حاصل اِعمال فیلتر بر سیگنال ورودی به‌صورت زیر محاسبه می‌شود.

$$y[n] = x[n] * h[n] = \sum_{l=0}^{L-1} x[l].h[n-l] \qquad (۱)$$

که در آن L، طول فیلتر است.

اخیراً برخی از روش‌های جدیدتر، محدودیت‌هایی را روی فیلترهای لایه اول اِعمال می‌کنند که منجر به تنوع بیشتر روش‌هایی شده است که از سیگنال خام بعنوان ورودی استفاده می‌کنند. به‌عنوان مثال، مدل سینک نت محدود به یادگیری فیلتر سینک در لایه اول می شود. از آنجاکه این فیلترها، معادله حوزه زمان ساده‌ای دارند، می‌توانند با تعداد پارامترهای بسیار کم‌تر بیان شوند. بنابراین استفاده از این فیلترها، تعداد پارامترهای قابل تعلیم مدل را کاهش داده و سرعت همگرایی آن را افزایش می‌دهد.

۱.۳. انواع فیلترهای پارامتری

در این بخش فیلترهایی که در کاربردهای مختلف پردازش گفتار و معماری‌های ژرف با تعداد پارامتر محدود مورد استفاده قرار گرفته‌اند، معرفی می‌شوند.

۱.۱.۳. فیلتر میان‌گذر مستطیلی

اندازه فیلتر میان‌گذر مستطیلی[۲۴] را در حوزه فرکانس می‌توان به‌صورت تفاضل دو فیلتر پایین‌گذر[۲۵] در نظر گرفت.

$$H_{\text{rect}}(f) = \text{rect}\left(\frac{f}{2f_2}\right) - \text{rect}\left(\frac{f}{2f_1}\right) \qquad (۲)$$

که در آن f_1 و f_2 فرکانس‌های قطع بالا و پایین نرمال‌شده بوده و در طی فرایند آموزش یاد گرفته می‌شوند. rect(.) نیز تابع مستطیلی[۲۶] نمایانگر اندازه طیف است. معادل حوزه زمان این تابع، تابع سینک است و در نتیجه می توان پاسخ ضربه این فیلتر را به شکل زیر نوشت.

$$h_{\text{rect}}[n] = 2f_2\text{Sinc}(2\pi f_2 n) - 2f_1\text{Sinc}(2\pi f_1 n) \qquad (۳)$$

که در آن تابع سینک به‌صورت $\text{Sinc}(x) = {\sin(x)}/{x}$ تعریف می‌شود. بنابراین، فیلتر حاصل یک فیلتر متقارن است و در نتیجه باعث اعوجاج فاز[۲۷] نمی‌شود. برای اینکه اطمینان داشته‌باشیم که $f_1 \geq 0$ و $f_2 \geq f_1$ است در واقع مقادیر زیر به‌عنوان f_1 و f_2 به معادله فوق اعمال می‌شوند.

$$f_1^{\text{abs}} = |f_1| \qquad (۴)$$

$$f_2^{\text{abs}} = f_1 + |f_2 - f_1|$$

مقادیر فرکانس‌های قطع می‌توانند به‌صورت تصادفی در بازه $[0, f_s/2]$ مقداردهی اولیه شوند که در آن f_s فرکانس نمونه‌برداری سیگنال ورودی است. بسته به کاربرد، می‌توان فیلترها را با مقادیر فرکانس قطع بانک فیلتر در مقیاس مل نیز مقداردهی اولیه کرد. این نوع مقداردهی اولیه باعث می‌شود فیلترهای بیشتری در فرکانس‌های پایین عمل کنند. بعلاوه، بهره[۲۸] هر فیلتر یاد گرفته نمی‌شود بلکه توسط لایه‌های بعد تعیین می‌شود که هر فیلتر چقدر در خروجی لایه بعد مؤثر است.

برای اینکه بتوان از این فیلتر در معماری پیچشی استفاده کرد، باید آن را بُرش[۲۹] داد. این عمل باعث می‌شود که فیلتر حاصل، تخمینی از فیلتر ایده آل بوده و دارای ریپل[۳۰] در باند عبور[۳۱] و تضعیف[۳۲] در باند توقف[۳۳] باشد. یکی از راه‌حل‌های مرسوم برای کم کردن این اثر، پنجره‌گذاری[۳۴] [۱۵] است. پنجره‌گذاری با ضرب تابع بریده شده پاسخ ضربه h در تابع پنجره w به دست می‌آید و باعث می‌شود ناپیوستگی‌های انتهایی h نرم‌تر شوند.

$$h_w[n] = h[n].w[n] \qquad (۵)$$

یکی از پرکاربردترین پنجره‌ها، پنجره همینگ[۳۵] است که در رابطه ۶ تعریف می‌شود. می توان بسته به کاربرد از پنجره‌های دیگری مانند هنینگ[۳۶]، بلکمن[۳۷] و کایزر[۳۸] نیز استفاده کرد.

$$w_{\text{hamming}}[n] = 0.54 - 0.46\,.\cos\left(\frac{2\pi n}{L}\right) \qquad (۶)$$

براساس [۱۶]، می‌توان فیلتر مستطیلی را به‌صورت حاصلضرب یک کرنل در یک حامل[۳۹] نوشت. با این روش، هر فیلتر را می‌توان با استفاده از دو پارامتر فرکانس مرکزی (f_c) و پهنای باند (B) مشخص نمود که در آن $f_c = (f_1 + f_2)/2$ و $B = f_2 - f_1$ است. معادله زیر نمایش پاسخ ضربه فیلتر مستطیلی را به این صورت جدید نشان می‌دهد.

$$h_{\text{rect}}[n] = 2B\,\text{Sinc}(Bn)\cos(2\pi f_c n) \qquad (۷)$$

سایر فیلترهایی که در ادامه معرفی می‌شوند را می توان به همین ترتیبی که برای فیلتر مستطیلی توضیح داده شد، در مدل‌سازی با معماری پیچشی لحاظ نمود.

۲.۱.۳. فیلتر مثلثی

شکل فیلترهای شنیداری را می‌توان با فیلترهای مثلثی نیز تقریب زد [۱۷]. در کاربردهای سنتی، بانک فیلتر متداول در مقیاس مل معمولاً با استفاده از این نوع فیلتر محاسبه می‌شود. پاسخ ضربه فیلتر مثلثی در حوزه زمان تابع Sinc^2 است که معادله آن به‌صورت زیر می‌باشد.

$$h_{\mathrm{tri}}[n] = \mathrm{Sinc}^2(Bn)\cos(2\pi f_c n) \qquad (۸)$$

۳.۱.۳. فیلتر گاماتون[۴۰]

فیلتر گاماتون که در [۱۸] معرفی شده است، در مدل‌سازی شنیداری بسیار رایج است. این فیلترها تطابق خوبی با توابع رِوکُر[۴۱] دارد. این توابع از مطالعات فیزیولوژیک انجام شده در مورد حلزونی گوش گربه‌ها حاصل شده‌اند. پاسخ ضربه فیلترهای گاماتون بسیار ساده است، حال‌آنکه پاسخ فرکانسی آن پیچیده است. معادله زیر پاسخ ضربه این نوع فیلتر را بیان می‌کند.

$$h_{\mathrm{gammatone}}[n] = n^{N-1}\,\mathrm{e}^{-2\pi Bn}\cos(2\pi f_c n) \qquad (۹)$$

که در آن N درجه فیلتر است. معمولا N را برابر ۴ در نظر می‌گیرند چراکه این مقدار به‌خوبی با فیلترهای حلزونی گوش تطابق دارد [۱۶].

۴.۱.۳. فیلتر گاماچیپ[۴۲]

فیلتر گاماتون نمی‌تواند به طور کامل مطابق با مشخصه‌های فیلتری فرکانس غشای حلزونی گوش باشد. بانک فیلتر گاماچیپ جایگزینی برای این فیلترها است. این بانک فیلتر در عین داشتن مزایای بانک فیلتر گاماتون، معایب مربوط به تقارن و وابستگی به شدت را ندارد. به همین دلیل، این نوع فیلتر می‌تواند مشخصه‌های روانی-فیزیولوژیکی فیلتری سامانه شنیداری انسان را در محدوده بزرگی از فرکانس‌های مرکزی تولید کند [۱۹]. پاسخ ضربه این فیلتر با استفاده از فرمول زیر بیان می‌شود.

$$h_{\mathrm{gammachirp}}[n] = n^{N-1}\,\mathrm{e}^{-2\pi Bn}\cos(2\pi f_c n + c.\ln(n) + \varphi) \qquad (۱۰)$$

که در آن N درجه فیلتر، c نرخ افزایش فرکانس[۴۳] و φ فاز اصلی است.

۵.۱.۳. فیلتر گاوسی

فیلتر گوسی را می‌توان به‌عنوان یک فیلتر گاماتون درجه بالا در نظر گرفت. پاسخ ضربه این نوع فیلتر به‌صورت زیر است.

$$h_{\mathrm{gauss}}[n] = \mathrm{e}^{-n^2/2\sigma^2}\cos(2\pi f_c n) \qquad (۱۱)$$

با این فرض که B پهنای باند 3dB است، σ را می‌توان برابر $\sqrt{\log 2}/2\pi B$ در نظر گرفت.

۶.۱.۳. فیلتر گابور[۴۴]

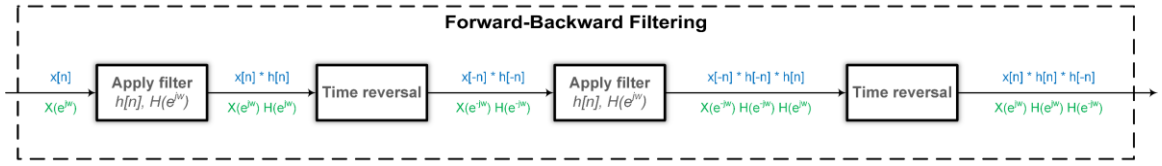
فیلتر گابور که از مادوله کردن یک کرنل گاوسی با یک سیگنال سینوسی به دست می‌آید، در مرجع [۲۰] در معماری که به LEAF [۴۵] موسوم است، مورد استفاده قرار گرفته است. فیلتر گابور مختلط که در [۲۱] استفاده شده است، تفکیک‌پذیری[۴۶] زمان – فرکانس بهتری فراهم می‌آورد. معادله زیر پاسخ ضربه فیلتر گابور را نمایش می‌دهد.

$$h_{\mathrm{gabor}}[n] = \frac{1}{\sqrt{2\pi}B}\mathrm{e}^{-\frac{n^2}{2B^2}}\mathrm{e}^{j2\pi f_c n} \qquad (۱۲)$$

که در آن $A = \sqrt{\frac{3\ln 10}{10}}, B = \frac{A}{\pi(f_2 - f_1)}$ است.

۷.۱.۳. پنجره پارزن[۴۷]

در [۲۲] از تابع پنجره مربعی اپانچنیکوف[۴۸] که با یک سیگنال کسینوسی مادوله می‌شود، به‌عنوان فیلترهایی



شکل ۳. عملیات لازم برای بدست آوردن سیگنال با فاز اصلاح شده پس از اعمال فیلتر **IIR** بریده شده [۵]

که در لایه اول یاد گرفته می‌شوند، استفاده شده است. معادله زیر فرمول این فیلتر که به پنجره پارزن موسوم است، را نشان می‌دهد.

$$h_{\text{parzen}}[n] = 2\pi f_c n\, k[n] \qquad (۱۳)$$

$$k[n] = \begin{cases} (1 - Bn^2)^2, |n| \le {}^{1}/_{\sqrt{B}} \\ 0, \quad \text{otherwise} \end{cases} \qquad (۱۴)$$

۸.۱.۳. فیلتر آبشاری

اگر چه بیشتر آنالیزهای فرکانسی بر اساس فیلترهای خطی نامتغیر با زمان (LTI) انجام می‌شود، اما سیستم شنیداری به‌صورت خطی پاسخ نمی‌دهد. رابطه غیرخطی بین ورودی و خروجی در زمانی که ورودی دو برابر می‌شود یا بهره پیدا می‌کند و یا پهنای باند به‌عنوان تابعی از شدت صدا تغییر پیدا می‌کند، گواه این مدعاست. آبشاری کردن چند فیلتر خطی راهی برای پیچیده‌تر کردن فیلترهای ساده است. در آزمایشاتی که در [۲۳] انجام شده، از فیلتر آبشاری حاصل از پیچش دو فیلتر گوسی استفاده شده است. معادله زیر، پاسخ ضربه این فیلتر را نمایش می‌دهد.

$$h_{\text{cascade}}[n] = h_{\text{gauss},1}[2n] * h_{\text{gauss},2}[2n] \qquad (۱۵)$$

۹.۱.۳. فیلتر IIR تغییریافته

نمایش صفر-قطب فیلترهای IIR[۴۹] کسری بسیار کارامد است. ازآنجاکه پاسخ ضربه فیلترهای IIR نامتناهی است، عملاً امکان استفاده از آن‌ها به‌عنوان یک فیلتر با طول محدود در شبکه‌های عصبی امکان‌پذیر نیست. بعلاوه خطی بودن فاز فیلترهای IIR تضمین شده نیست. برای حل این مشکلات در [۵] فیلتر IIR بریده[۵۰] شده (h_{TIIR}) و سپس با استفاده از روش فیلتری پیشرو – پسرو[۵۱] (h_{IIRI})، فاز آن اصلاح می‌شود (شکل ۳).

$$h_{\text{TIIR}}[n] = \frac{\sin(2\pi(n+1)f_c)}{\sin 2\pi f_c} \mathrm{e}^{-Bn}\, u[n] \qquad (۱۶)$$

$$h_{\text{IIRI}}[n] = h_{\text{TIIR}}[n] * h_{\text{TIIR}}[-n] \qquad (۱۷)$$

۱۰.۱.۳. فیلتر تحلیلی[۵۲]

یکی از ویژگی‌های مطلوب بازنمایی زمان – فرکانس که فیلترهای تحلیلی نیز آن را دارند، آن است که نسبت به تاخیرهای کمی که در حوزه زمان وجود دارد، مقاوم هستند. مدول پیچش بین یک سیگنال حقیقی و فیلتر تحلیلی، برابر با پوش[۵۳] آن سیگنال در باند فرکانسی است که فیلتر عبور می‌دهد. فیلترهای تبدیل فوریه زمان کوتاه[۵۴] (STFT) نمونه‌هایی از فیلترهای تحلیلی هستند که اندازه STFT، بازنمایی مقاوم به جابجایی[۵۵] آن است [۲۴]. برای هر فیلتر حقیقی مانند $u(t)$ می‌توان یک فیلتر تحلیلی با استفاده از فرمول زیر بدست آورد.

$$u_{\text{analytic}}(t) = u(t) + j\mathcal{H}[u(t)] \qquad (۱۸)$$

که در آن $\mathcal{H}$ تبدیل هیلبرت است و فاز هر مولفه فرکانسی مثبت را به اندازه $\pi/2$- انتقال می‌دهد. فیلترهای تجزیه تحلیلی پارامتری[۵۶] را به‌صورت زیر تعریف می‌کنند.

$$\begin{aligned} h_{\text{analytic}}[n] = {} & 2B\ \text{Sinc}(Bn)(\cos(2\pi f_c n) \\ & - j\sin(2\pi f_c n)) \\ = {} & 2B\ \text{Sinc}(Bn)\ \mathrm{e}^{-2j\pi f_c n} \end{aligned} \quad (۱۹)$$

که در آن تحلیلی بوده هر فیلتر به علت درستی تساوی زیر است:

$$\Im\left(h_{\text{analytic}}[n]\right) = \mathcal{H}[\Re\left(h_{\text{analytic}}[n]\right)] \quad (۲۰)$$

خانواده فیلترهای سنتز متناظر با فیلترهای فوق به صورت زیر تعریف می‌شوند:

$$h_{\text{synthesis}}[n] = 2AB\ \text{Sinc}(Bn)\mathrm{e}^{2j\pi f_c n} \quad (۲۱)$$

که در آن A پارامتر بهره است و برای بهبود سنتز یاد گرفته می‌شود.

۱۱.۱.۳. فیلتر شخصی شده[۵۷] (PF)

مرجع [۲۵] سعی کرده شکل حوزه فرکانس فیلترها را با تعریف تعدادی نقطه شکست[۵۸] یاد بگیرد. این نقاط، قطعه خط‌هایی را ایجاد می‌کنند که معادله هر قطعه به‌صورت زیر تعریف می‌شود.

$$\begin{aligned} H_{\text{PF}}[f, f_k, f_{k+1}] = {} & \frac{h_{k+1} - h_k}{f_{k+1} - f_k}\,(f - f_k) \\ & + h_k \end{aligned} \quad (۲۲)$$

در این معادله، f_k و f_{k+1} فرکانس‌های بالا و پایین هر قطعه را تعیین کرده و h_k و h_{k+1} نشان دهنده اندازه آن‌ها هستند. با پشت سر هم قرار دادن این خطوط، کل باند فرکانسی پوشش داده می‌شود. با اعمال تبدیل فوریه معکوس، معادله حوزه زمان هر قطعه به‌صورت زیر بیان می‌شود.

$$\begin{aligned} & h_{\text{PF}}[n, f_k, f_{k+1}] \\ & = \frac{\Delta(\cos(2\pi f_{k+1} n) - \cos(2\pi f_k n))}{4\pi^2 n^2} \\ & - \frac{h_{k+1}\sin(2\pi f_{k+1} n) - h_k \sin(2\pi f_k n)}{2\pi n} \end{aligned} \quad (۲۳)$$

که در آن $\Delta = \frac{h_{k+1} - h_k}{f_{k+1} - f_k}$ است.

۲.۳. مقایسه فیلترهای پارامتری

در این بخش فیلترهای معرفی شده در قسمت قبل مقایسه و نقاط قوت و ضعف هر کدام بیان می‌شوند. خلاصه این بحث در جدول ۱ آمده است.

فیلتر میان‌گذر مستطیلی، ساده بوده و محاسبات کمی دارد. بعلاوه، به‌دلیل داشتن فاز خطی، باعث اعوجاج در فاز نمی‌شود. از سوی دیگر، شکل پاسخ فرکانسی این فیلتر دارای قطع‌های تیز[۵۹] بین فرکانس‌های باند عبور و باند توقف است. همچنین برای رسیدن به پاسخ فرکانسی مطلوب، بایستی طول فیلتر بزرگ‌تر در نظر گرفته شود که باعث مصرف بیشتر حافظه خواهد شد.

فیلتر مثلثی نیز ساده بوده، فاز خطی داشته و به خاطر گذار نرم از باند عبور به باند توقف، نشت طیفی کم‌تری دارد. تفکیک‌پذیری فرکانسی پایین این فیلتر در کاربردهایی که نیاز به تمایز فرکانسی دقیق دارند، ممکن است مناسب نباشد. بعلاوه لُب اصلی این فیلتر نیز می‌تواند عریض شود و در نتیجه آن را برای کاربردهایی که به قطع تیز نیاز دارند، نامناسب سازد.

فیلتر گاماتون در عین سادگی و هزینه محاسباتی پایین، می‌تواند مشخصه‌های غشای بازیلار حلزونی گوش را شبیه‌سازی کند. تداخل مؤلفه‌های فرکانسی در فیلتر گاماچیپ نسبت به فیلتر گاماتون، کمتر است اما این فیلتر نسبت به مقادیر پارامترهای آن حساس‌تر بوده و این امر می‌تواند منجر به بیش برازش مدل یادگیری شود. این فیلتر قابلیت انتخاب فرکانس[۶۰] سیستم شنیداری انسان را شبیه‌سازی کرده و تفکیک‌پذیری زمانی خوبی نیز فراهم می‌آورد. در نتیجه

برای کاربردهایی که اطلاعات زمانی در آن مهم است، مناسب می‌باشد.

فیلتر گاوسی نیز محاسبات ساده‌ای داشته و فاز خطی دارد. بعلاوه، ویژگی‌های مهم سیگنال گفتار مانند سازه‌ها[۶۱] و گام[۶۲] را حفظ می‌کند. اما خاصیت نرم کنندگی[۶۳] این فیلتر می‌تواند باعث حذف جزئیات شده و در نتیجه کیفیت گفتار را تحت تأثیر قرار دهد. بعلاوه، مدل می‌تواند نسبت به انتخاب پارامترهای این فیلتر، به‌خصوص انحراف معیار، حساس شده و مستعد بیش برازش شود. پنجره پارزن نیز همانند فیلتر گاوسی برای حذف نویز مناسب است و در نتیجه برای کاربردهایی که نویز پیش‌زمینه زیادی دارند، مناسب خواهد بود.

جدول ۱. فیلترهای پارامتری و معماری‌های مورد استفاده در کاربردهای مختلف پردازش گفتار

فیلتر	نقاط قوت	نقاط ضعف
میان‌گذر مستطیلی	– سادگی (تعداد پارامتر کم) – فاز خطی – هزینه محاسباتی پایین	– پاسخ فرکانسی نامناسب – نیاز به طول فیلتر بزرگ‌تر در مقایسه با فیلترهای IIR
مثلثی	– سادگی – نشت طیفی کمتر – فاز خطی – هزینه محاسباتی پایین	– تفکیک‌پذیری فرکانسی پایین – پاسخ فرکانسی نامناسب برای برخی از کاربردها
گاماتون	– تفکیک‌پذیری فرکانسی مناسب – شبیه‌سازی مشخصه‌های غشای بازیلار – هزینه محاسباتی پایین	– امکان تداخل بین مولفه‌های فرکانسی
گاماچیپ	– مدل‌سازی قابلیت انتخاب فرکانس سیستم شنیداری انسان – تفکیک‌پذیری زمانی مناسب – تداخل کمتر مولفه‌های فرکانسی	– محاسبات بیشتر – حساسیت بالا به مقادیر پارامترها و در نتیجه امکان بیش برازش مدل
گاوسی	– هزینه محاسباتی پایین – نگهداشت ویژگی‌های مهم گفتار مانند سازه‌ها و گام – فاز خطی	– حذف جزئیات – حساسیت به پارامترها
گابور	– تحلیل سیگنال در هر دو حوزه زمان و فرکانس به طور همزمان – تحلیل چند تفکیکی	– حساسیت به نویز – محاسبات بیشتر
پنجره پارزن	– حذف نویز	– حساسیت به مقدار پارامتر پهنای باند
آبشاری	– انعطاف پذیری در تولید فیلترهای دلخواه	– محاسبات بیشتر – پیچیدگی مدل – امکان به هم ریختن فاز سیگنال
IIR تغییریافته	– فاز خطی – تفسیرپذیری نمایش صفر و قطب	– محاسبات بیشتر
تحلیلی	– استخراج پوش سیگنال – انعطاف پذیری	– محاسبات بیشتر
شخصی شده	– انعطاف بیشتر در شکل پاسخ فرکانسی	– محاسبات بیشتر

بعلاوه این فیلتر نیز نسبت به پارامتر پهنای باند حساسیت داشته و می‌تواند باعث بیش برازش مدل شود.

با استفاده از فیلتر گابور می‌توان سیگنال را در هر دو حوزه زمان و فرکانس به‌طور همزمان تحلیل کرد. این خاصیت برای در نظر گرفتن ویژگی‌های گذرا[۶۴] مانند واج‌ها[۶۵] مناسب است. اما از سوی دیگر محاسبات بیشتری داشته و نسبت به نویز حساس است.

استفاده از چند فیلتر به‌صورت سِری (فیلتر آبشاری)، می‌تواند گذار بین باند عبور و توقف را تیزتر کند. بعلاوه می‌تواند انعطاف‌پذیری مدل را در تولید فیلترهای دلخواه با ترکیب فیلترهای بالاگذر و پایین‌گذر مناسب، بالاتر برد. اما در مقابل، استفاده از این نوع فیلتر، پیچیدگی مدل را بالاتر برده و محاسبات آن را سنگین‌تر می‌کند. همچنین می‌تواند باعث به‌هم‌ریختگی فاز سیگنال شود. در فیلتر IIR تغییریافته برای داشتن فاز خطی، از روش فیلتر پیشرو – پسرو استفاده شده است. این امر محاسبات این فیلتر را سنگین‌تر می‌کند. نمایش صفر و قطب این نوع فیلتر، ابزار مناسبی برای تفسیر و حتی بهینه‌سازی تعداد فیلترهای بانک فیلتر یادگرفته شده خواهد بود.

توانایی فیلترهای تحلیلی در استخراج پوش سیگنال گفتار، آن‌ها را برای کاربردهایی مانند بهسازی گفتار و تشخیص فعالیت صوتی مناسب می‌سازد. این فیلتر نیاز به محاسبات بیشتری دارد. در فیلترهای شخصی شده، تمرکز بر ایجاد فیلترهایی با شکل پاسخ فرکانسی دلخواه است اما برای به دست آوردن پاسخ ضربه باید تبدیل فوریه معکوس، اعمال شود که باعث افزایش هزینه محاسباتی می‌شود.

۴. معماری‌های دارای فیلتر پارامتری

در این بخش، مهم‌ترین معماری‌هایی که در کاربردهای مختلف از فیلترهای معرفی شده در قسمت قبل استفاده کرده‌اند، توضیح داده می‌شوند.

۱.۴. مدل‌های مبتنی بر شبکه‌های عصبی پیچشی

مدل سینک‌نت، اولین معماری است که از فیلترهای معنی‌دار در معماری ژرف برای کاربردهای پردازش گفتار استفاده کرده است. در این مدل، لایه اول پیچشی بجای فیلترهای دلخواه، فیلترهای میان‌گذر مستطیلی یاد می‌گیرد. این معماری شامل سه لایه پیچشی و سه لایه تمام متصل[۶۶] (FC) است. پس از هر لایه پیچشی یک لایه تجمیع[۶۷] به طول ۳، یک لایه نرمال‌سازی و تابع فعالیت Leaky ReLU مورد استفاده قرار می‌گیرد. در لایه اول پیچشی ۸۰ فیلتر مستطیلی با تعداد پارامتر $L=257$ یاد گرفته می‌شود. در هریک از دو لایه بعدی پیچشی ۶۰ فیلتر با تعداد پارامتر ۵ یاد گرفته می‌شوند. تعداد نورون‌ها در همه لایه‌های تمام متصل ۲۰۲۴ در نظر گرفته شده است.

در [۶] اثر استفاده از فیلترهای مختلفی همچون مثلثی، گاوسی، گاماتون و آبشاری در دقت این معماری در کاربرد تشخیص گوینده مورد بررسی قرار گرفته و با هدف ارائه توضیح رفتار مدل، فیلترهای یادگرفته شده از دیدگاه‌های مختلف تجزیه و تحلیل شده‌اند. در [۲۵, ۲۶] به ترتیب فیلترهای IIRI و PF برای استفاده در این معماری معرفی شده‌اند. در سال ۲۰۲۳ مرجع [۲۳] با بهره‌گیری از این فیلترها، روشی را برای کاهش تعداد فیلترهای لایه اول پیچشی برای تشخیص جنسیت گوینده ارائه می‌دهد.

در سال‌های اخیر، استفاده از سیگنال خام به‌عنوان ورودی یک سیستم بهبود کیفیت گفتار، بیشتر مورد توجه قرار گرفته است. دلیل این امر آن است که سیگنال خام، مشکل تخمین نادرست فاز را در مقایسه با دیگر انواع ورودی ندارد. مرجع [۲۷] یک روش بهبود کیفیت گفتار چندکاناله که از یک شبکه پیچشی کامل[۶۸] (FCN) با لایه‌های پیچشی سینک و منبسط شونده[۶۹] ارائه داده است. در [۲۸] از معماری Wav-UNet [۲۹] با فیلترهای گابور برای کاربرد بهبود کیفیت گفتار استفاده شده است.

RawNet2 [۳۰] که به‌عنوان بهبودی بر RawNet [۳۱]، ارائه شده، معماری برای تصدیق هویت گوینده است که در لایه اول آن از سینک استفاده شده است. لایه‌های بعدی به ترتیب ۲ و ۴ بلوک باقیمانده با فیلترهای پیچشی با اندازه‌های مختلف، واحد بازگشتی دروازه‌دار[۷۰] (GRU) و لایه تمام متصل هستند. در این معماری از مقیاس‌بندی نقشه ویژگی[۷۱] (FMS) استفاده شده است که در آن یک بردار مقیاس با تابع غیرخطی سیگموید[۷۲] برای مقیاس بندی نقشه‌های ویژگی در نظر گرفته می‌شود. معماری این شبکه در جدول ۲ آمده است.

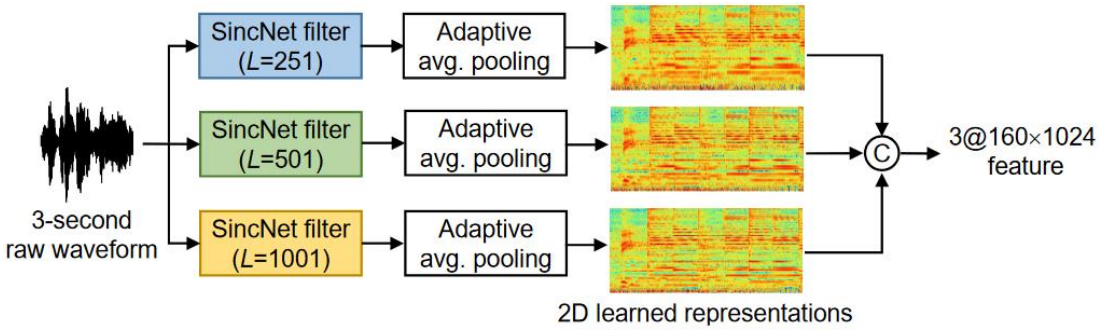
اندازه کرنل فیلترها در لایه‌های پیچشی استاندارد برابر یکدیگر در نظر گرفته می‌شود. به همین خاطر یادگیری اطلاعات فرکانسی بالا و پایین از یک سیگنال پهن باند[۷۳]، چالش‌برانگیز می‌شود. برای حل این مشکل، یک لایه پیچشی به چند لایه موازی با اندازه‌های مختلف شکسته شده و بدین ترتیب بازنمایی سطح پایین بهتری از سیگنال خام ورودی به دست می‌آید. از این روش که در مراجع [۳۳, ۳۴] استفاده شده است، به‌عنوان فیلترهای چند مقیاسی[۷۴] یاد می‌شود. بدین ترتیب، پارامترهای مختلف هر لایه موازی پیچشی شامل تعداد فیلترها، اندازه کرنل و اندازه گام[۷۵] به‌صورت جداگانه تعیین شده و فیلترهای هر مقیاس به مؤلفه‌های فرکانسی مختلف پاسخ می‌دهند.

جدول ۲. معماری شبکه **RawNet2** [۳۲]

Layer	Input:59049 samples	Output shape
Sinc -conv	Sinc(251,1,128) MaxPool(3) BN LeakyReLU	(19683, 128)
Res block	{ BN LeakyReLU Conv(3,1,128) BN LeakyReLU Conv(3,1,128) MaxPool(3) FMS } ×2	(2187, 128)
Res block	{ BN LeakyReLU Conv(3,1,256) BN LeakyReLU Conv(3,1,256) MaxPool(3) FMS } ×4	(27, 256)
GRU	GRU(1024)	(1024,)
Speaker embedding	FC(1024)	(1024,)

در [۳۴]، برای دسته‌بندی ژانر موسیقی، سیگنال خام در ابتدا به سه لایه فیلتری سینک داده شده و سپس حاصل اعمال فیلترها به سیگنال ورودی، در قالب یک تصویر سه کاناله به شبکه ResNet داده می‌شود. سه مجموعه فیلتر سینک با ۱۶۰ کرنل با طول‌های ۲۵۱، ۵۰۱ و ۱۰۰۱، تفکیک‌پذیری‌های فرکانسی متفاوتی از سیگنال ورودی را ایجاد می‌کنند. از لایه تجمیع میانگین تطبیقی[۷۶] برای داشتن خروجی با ابعاد مشابه در هر یک از سه لایه استفاده شده است. شکل ۴، این بخش از معماری را نمایش می‌دهد.

در [۳۵] نیز از فیلترهای سینک چند مقیاسی برای تشخیص احساس در موسیقی استفاده شده است. در اینجا، از بازنمایی استخراج شده از شبکه ResNet، تعدادی ویژگی محلی و عمومی استخراج می‌شود که

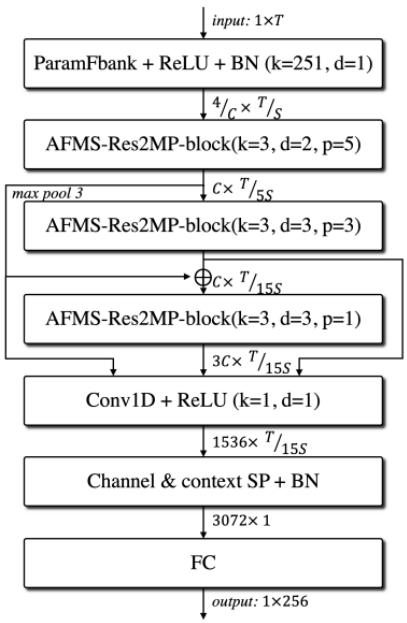


شکل ۴. نحوه تشکیل تصویر سه کاناله با استفاده از فیلترهای سینک چند مقیاسی [۳۴]

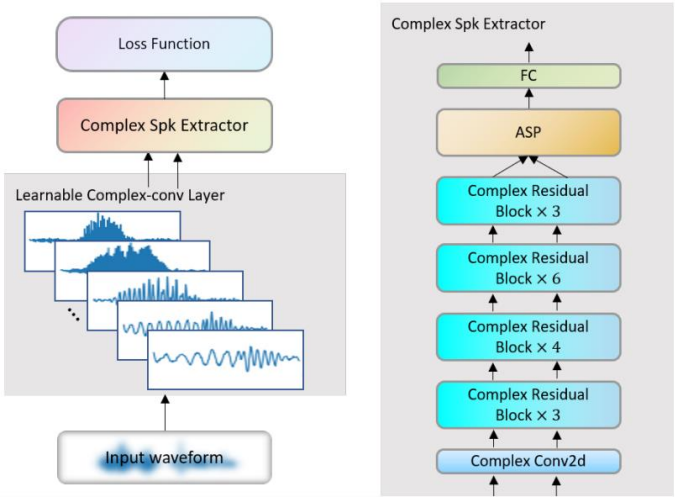
در یک ماژول ادغام، ترکیب شده و بر پایه آن نمونه ورودی دسته‌بندی می‌شود. مرجع [۳۶] اثر مقداردهی اولیه فیلترهای گابور در معماری LEAF را بررسی می‌کند. نتایج ارائه شده در این مقاله نشان می‌دهند که اگرچه استفاده از فیلترهای معنی‌دار در لایه اول پیچشی، کارایی مدل را بهبود می‌دهند اما بانک فیلتر یادگرفته شده، حساسیت زیادی به نحوه مقداردهی اولیه فیلترها دارد. تغییرات ناچیز بانک فیلتر یاد گرفته شده نسبت به مقداردهی اولیه آن‌ها نشان می‌دهد که توسعه روش‌های دیگری برای یادگیری این بانک فیلتر می‌تواند کارایی کلی را بهبود بخشد.

در [۳۷] از یادگیری خود نظارتی مقابله‌ای[۷۷] برای تشخیص ناهنجاری‌های تنفسی مبتنی بر سینک‌نت استفاده شده است. در مرجع [۳۸] از سه مجموعه داده صوتی، لرزش و فراداده در کنار هم برای تشخیص عیب اتومبیل استفاده می‌کند. در این معماری از سینک‌نت برای استخراج ویژگی‌های آکوستیکی استفاده می‌شود.

در [۳۹, ۴۰] از تبدیل موجک به‌جای اعمال توابع سینک، برای شناسایی گوینده استفاده شده و لایه ScatNet نام گرفته است. مرجع [۴۱] با الهام از ECAPA-TDNN [۴۲] از فیلترهای Sinc و Gabor برای تصدیق گوینده استفاده می‌کند. RawNet3 [۲] معماری دیگری است که بر اساس ترکیب دو معماری RawNet2 و ECAPA-TDNN [۴۲] برای شناسایی گوینده ارائه شده است. شکل ۵ معماری این شبکه را نمایش می‌دهد. در لایه اول این معماری، یک بانک فیلتر تحلیلی با



شکل ۵. معماری **RawNet3** برای شناسایی گوینده [۲]



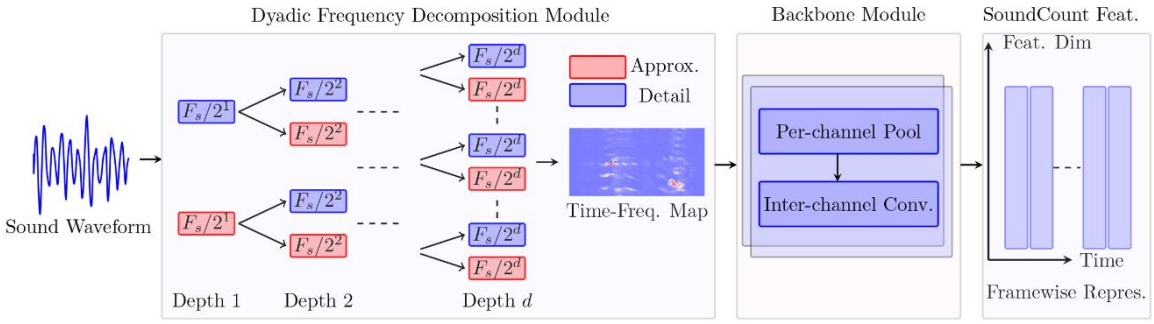
شکل ۶. معماری شبکه عصبی مختلط ارائه شده در [۳] برای تصدیق گوینده

مقادیر مختلط [۲۴] یاد گرفته می‌شود. مرجع [۲۱] با هدف بهره‌گیری کامل از تفکیک‌پذیری زمان-فرکانس سیگنال ورودی، از فیلترهای گابور مختلط در یک معماری شبکه عصبی مختلط ژرف برای کاربرد شناسایی گفتار استفاده کرده است. در [۳] فیلترهای STFT به فیلترهای نمایی مختلط تغییر پیدا کرده‌اند که فرکانس‌های آن با استفاده از یک شبکه عصبی مختلط ژرف یاد گرفته می‌شوند. شکل ۶ معماری این شبکه را نشان می‌دهد.

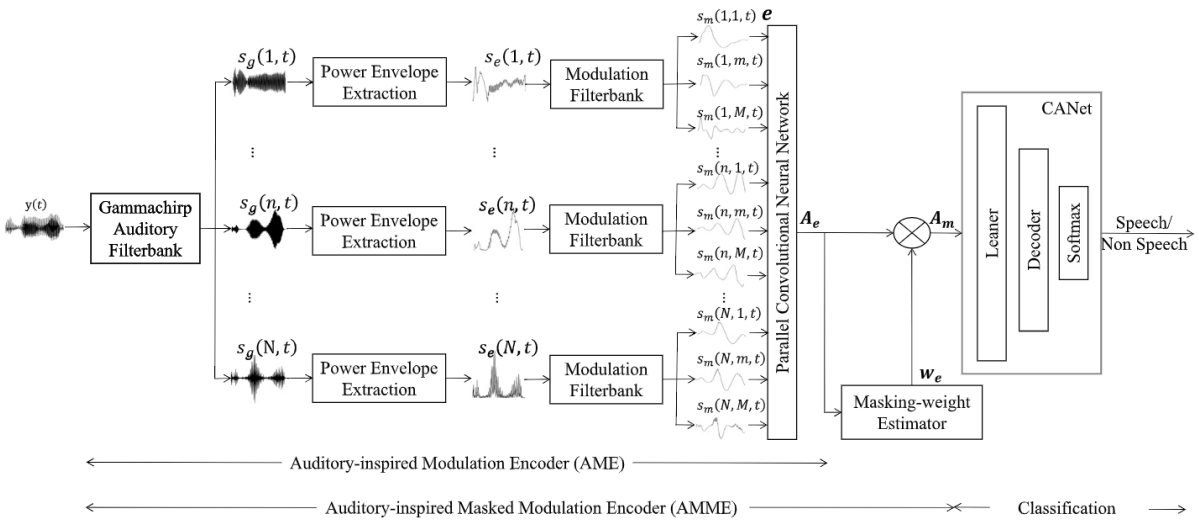
مرجع [۴۳] معماری به نام DyDecNet ارائه داده است که شامل یک ماژول تجزیه فرکانسی دوتایی است. در این ماژول مجموعه‌ای از بانک‌های فیلتر به‌صورت سلسله مراتبی یاد گرفته می‌شوند. در بانک فیلتر d اُم تعداد 2^d فیلتر Sinc وجود دارد. این بانک‌های فیلتر در یک ساختار آبشاری پشت سرهم قرار می‌گیرند و در نتیجه حوزه فرکانس سیگنال به‌طور پی‌درپی با پیشروی در عمق‌های بیشتر، با مضربی از ۲ تجزیه می‌شود. این ساختار، مشابه یک درخت دودویی کامل است که در آن با افزایش عمق درخت، تعداد فیلترها دو برابر و طول پاسخ فرکانسی نصف می‌شود و هر فرزند یکی از نیمه‌های بالا یا پایین پاسخ فرکانسی را تجزیه می‌کند. فیلترهایی که فرکانس‌های بالاتر را تجزیه می‌کنند، جزئیات[۷۸] و فیلترهایی که فرکانس‌های پایین را تجزیه می‌کنند، تخمین‌ها[۷۹] را کدگذاری می‌کنند. خروجی نهایی این ماژول، یک بازنمایی زمان فرکانس از سیگنال خام ورودی است که برای پردازش بیشتر به لایه‌های بعدی داده می‌شود. شکل ۷ معماری کلی ارائه شده در این تحقیق را نمایش می‌دهد.

۲.۴. مدل‌های مبتنی بر سازوکار توجه

مرجع [۱۹] یک معماری پیچشی مبتنی بر سازوکارتوجه برای تشخیص فعالیت صوتی را ارائه می‌دهد. در شکل ۸ نمای کلی این معماری نشان داده شده است. این معماری شامل ۳ مؤلفه اصلی AME [80]، AMME [۸۱] و CANet [۸۲] است و سعی دارد تمامی سامانه شنیداری انسان را تقلید کند.



شکل ۷. معماری ارائه شده در [۴۳] برای شمارش تعداد صداها



شکل ۸. معماری **AMME-CANet** برای تشخیص گفتار ارائه شده در [۱۹]

در این معماری، بلوک AME شامل ۴ مؤلفه اصلی بانک فیلتر گاماچیپ، استخراج پوش توان[۸۳]، بانک فیلتر مدولاسیون[۸۴] و شبکه عصبی پیچشی موازی[۸۵] است. به‌دلیل وجود تداخلات نویزی، ویژگی‌های به دست آمده از ماژول AME دسته بند مقاومی[۸۶] به دست نمی‌دهد. ازاین‌رو سعی می‌شود با شبیه‌سازی اثر ماسک گوش انسان در فرکانس‌های مختلف، یک ماژول تخمین زننده وزن ماسک[۸۷] به معماری اضافه شود. ماژول CANet شامل یک بلوک یادگیر[۸۸] و یک رمزگشا[۸۹] است که در بلوک یادگیر آن از لایه‌های پیچشی و مکانیزم توجه استفاده می‌شود.

در [۴۴] یک روش استخراج ویژگی الهام گرفته شده از سازوکار توجه برای کاربرد تصدیق گوینده ارائه شده است. ایده این تحقیق بر ایجاد تمایز بین گوینده‌ها بر اساس بیان عبارات کوتاهی مانند Hmmm که اطلاعات زبانی کمی دارند، استوار است. این عبارات کمتر تحت تأثیر اتفاقی بودن تلفظ نسبت به کلمات دیگر هستند. در اینجا از یک روش وزن‌دهی ویژگی تطبیقی جدید که از مکانیزم توجه الهام گرفته شده است، برای بهبود کیفیت ویژگی‌هایی که برای هر گوینده استخراج می‌شود، استفاده شده است.

در [۴۵] از فیلترهای گابور مختلط در یک معماری مبتنی بر توجه برای کاربرد تشخیص افسردگی استفاده

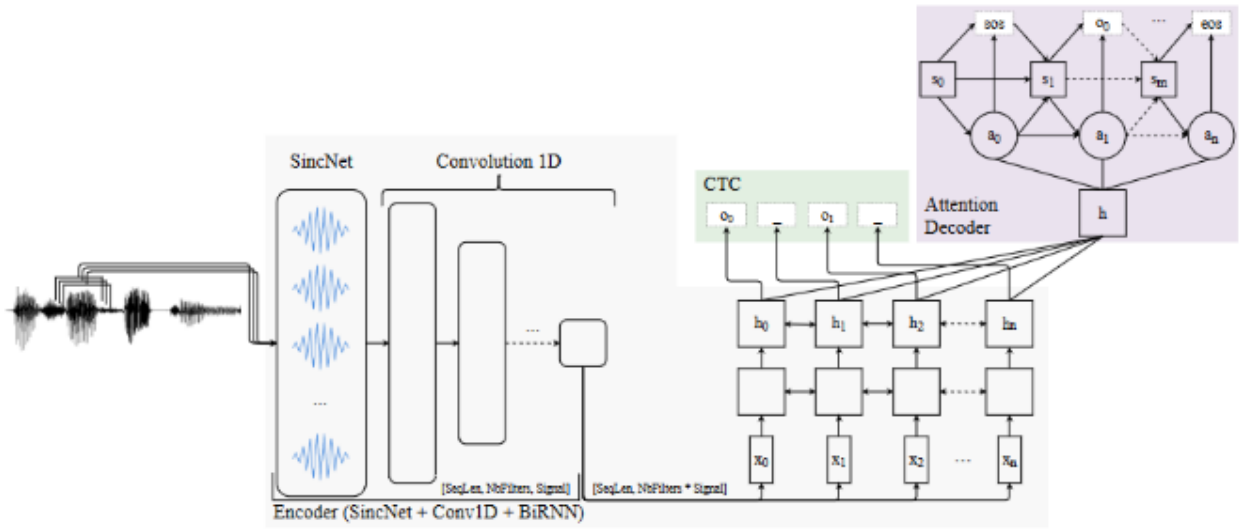
شده است. معماری ارائه شده در این تحقیق، شامل ۳ بخش اصلی است. در بخش اول، تعدادی فیلتر گابور مختلط یاد گرفته می‌شوند تا با استفاده از آن‌ها ویژگی‌های هر دو حوزه زمان و فرکانس استخراج شوند. در مرحله بعد، با بهره‌گیری از یک ماژول توجه چند مقیاسی طیفی، ویژگی‌های آکوستیکی که تمایز بیشتری بین دسته‌ها ایجاد می‌کنند، استخراج می‌شوند. در نهایت ماژول دسته‌بندی با استفاده از روش رأی‌گیری، برچسب را تعیین می‌کند.

۳.۴. مدل‌های مبتنی بر شبکه‌های عصبی بازگشتی

در [۴۶] از لایه Sinc در ترکیب با روش آموزش مبتنی بر CTC-attention یک معماری انتها به انتها برای شناسایی گفتار ارائه شده است. استفاده از شبکه‌های عصبی بازگشتی دوطرفه (BRNN) [۹۰] در این معماری و ترکیب آن با لایه Sinc عملکرد مدل را به‌طور قابل ملاحظه‌ای بهبود می‌دهد.

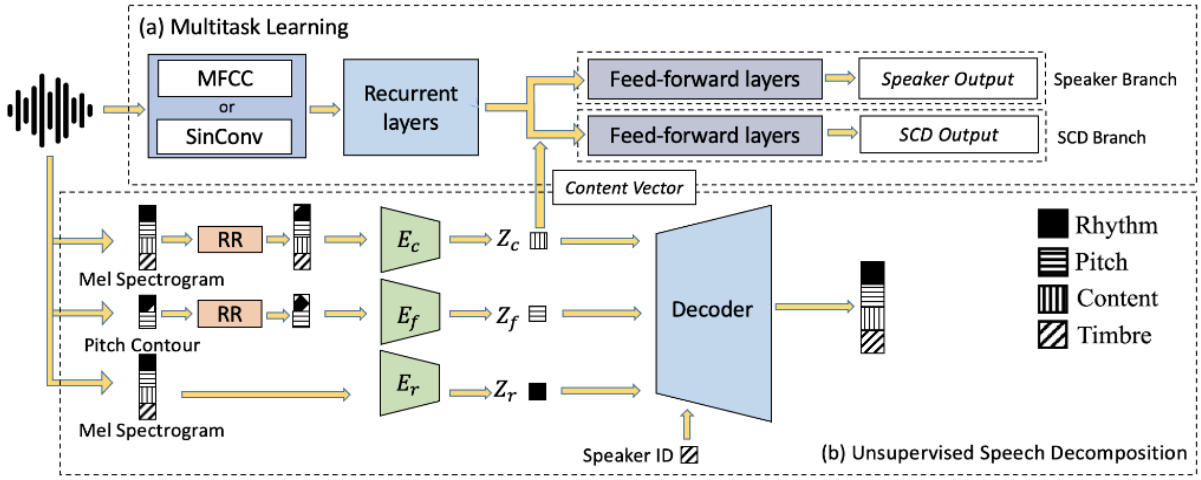
شکل ۹ معماری این مدل را نمایش می‌دهد. این معماری از سه مؤلفه اصلی تشکیل شده است. الف) بخش رمزگذار[۹۱] شامل چند لایه پیچشی است که اولین لایه آن از نوع سینک است. ویژگی‌های استخراج شده از این لایه‌ها سپس به یک شبکه عصبی بازگشتی دوطرفه داده می‌شود. ب) مؤلفه دوم، یک رمزگشای CTC است که برای هر گام زمانی یک توکن[۹۲] تولید می‌کند. ج) مؤلفه آخر یک رمزگشای مبتنی بر توجه[۹۳] است که بر اساس رشته توکن‌های رمزگشایی شده، نماد[۹۴] مناسب را در خروجی تولید می‌کند. مدل با استفاده از تابع زیان CTC-attention آموزش داده می‌شود.

در [۴۷] از یک معماری بازگشتی مبتنی بر LSTM[۹۵] برای تشخیص همزمان فعالیت صوتی و همپوشانی گفتار استفاده شده است. این معماری که در دو بخش از لایه‌های آن، خروجی‌های میانی گرفته می‌شود، یک ماژول extractor دارد که وظیفه آن استخراج ویژگی از سیگنال خام است. لایه اول این ماژول یک لایه سینک است.



شکل ۹. معماری ارائه شده در [۴۶]

مرجع [۴۸] از یک معماری یادگیری چند کاربردی[۹۶] استفاده می‌کند تا بتواند از اطلاعات محتوای گفتاری گوینده برای تشخیص تغییر گوینده در سطح قاب (فریم) استفاده کند. در این تحقیق از مجموعه ویژگی‌های MFCC در کنار ویژگی‌هایی که از لایه سینک استخراج می‌شود به‌عنوان ورودی دو لایه LSTM دوطرفه استفاده شده است. به‌طور موازی، با استفاده از ماژول تجزیه گفتار بدون نظارتی[۹۷] که در [۴۹] ارائه شده است، یک بردار محتوا[۹۸] استخراج می‌شود. ویژگی‌های استخراج شده از این دو ماژول به‌عنوان ورودی به دو شبکه پیش رو[۹۹] داده می‌شوند تا در نهایت بر اساس آن‌ها شخص گوینده و تغییر گوینده تشخیص داده شود. شکل ۱۰، معماری ارائه شده در این تحقیق را نمایش می‌دهد.



شکل ۱۰. معماری ارائه شده در [۴۸] برای تشخیص تغییر گوینده

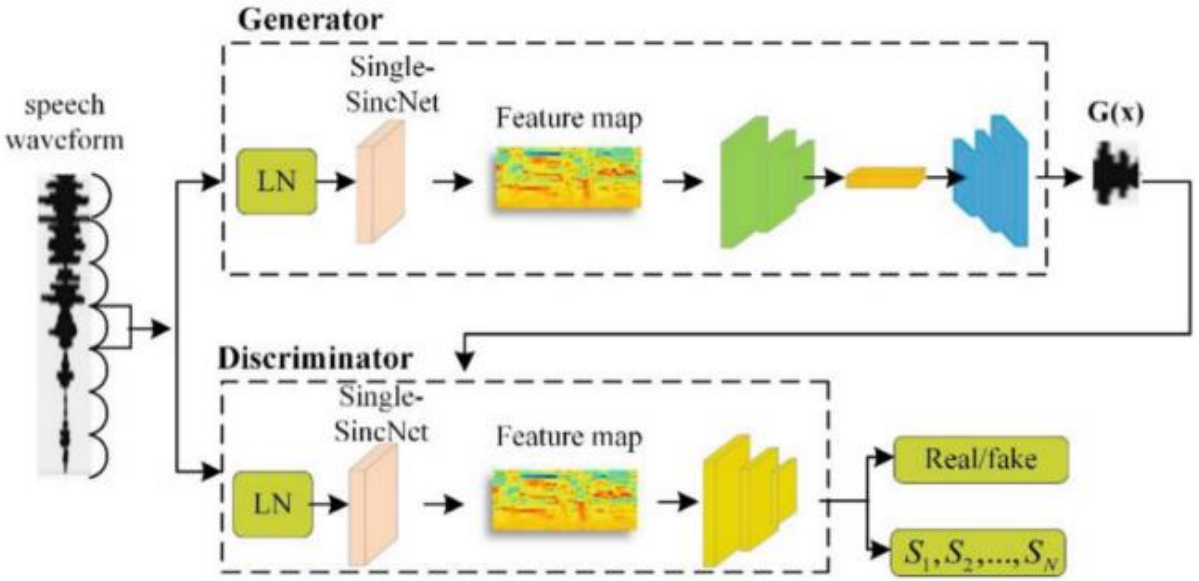
در [۵۰] برای کاربرد تشخیص گفتار فلجی[۱۰۰] از توالی از لایه‌های پیچشی، LiGRU[۱۰۱] و تمام متصل استفاده شده است. در این تحقیق از فیلترهای مستطیلی، مثلثی، گاوسی، گاماتون و پارزن در لایه اول پیچشی استفاده شده است. نتایج نشان داده‌اند که فیلتر گاوسی در این کاربرد، عملکرد بهتری داشته است. برای تشخیص اختلالات مرتبط با بیماری‌های عصبی از روی گفتار، در مرجع [۵۱] از یک معماری که حاصل ترکیب لایه‌های پیچشی، Bi-LSTM و مکانیزم توجه است، استفاده شده است.

۴.۴. مدل‌های مبتنی بر معماری‌های رقابتی

ترکیب فیلترهای پارامتری با معماری‌های رقابتی نیز مورد توجه قرار گرفته است. در این بخش، برخی از مهم‌ترین این مدل‌ها بررسی می‌شوند.

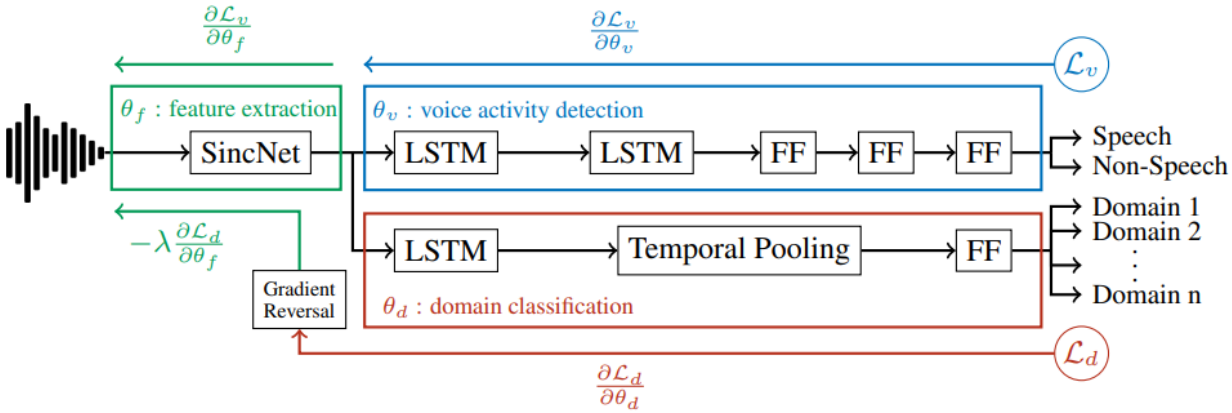
در [۵۲] از معماری DCGAN که ترکیبی از شبکه‌های عصبی پیچشی و شبکه‌های مولد رقابتی است، برای کاربرد تشخیص گوینده استفاده شده است. شکل ۱۱ معماری ارائه شده در این مرجع را نمایش می‌دهد. در ابتدای هر دو ماژول مولد و جداساز این معماری، یک لایه نرمال‌سازی (LN) وجود دارد که سیگنال ورودی را به لایه سینک می‌دهد. در ماژول مولد، برای افزایش نرخ نمونه‌برداری[۱۰۲]، تمام لایه‌های تجمیع با پیچشی ترانهاده[۱۰۳] جایگزین شده و در ماژول جداساز، تمام

لایه‌های تجمیع با افزودن گام[۱۰۴] به لایه‌های پیچشی جایگزین شده‌اند.



Discriminator *D*	Generator *G*
Input: X_{real},X_{fake}	Input:X_{real}
LN ([1, 3200])	LN ([1, 3200])
Single-SincNet Filter (K = 80, L = 251)	Single-SincNet Filter (K = 80, L = 251)
5 × 5 Conv2d. 64, BN, LR	15 × 1 Conv2d. 256, BN, LR
5 × 5 Conv2d. 128, BN, LR	5 × 1 Conv2d. 512, BN, LR
5 × 5 Conv2d. 256, BN, LR	5 × 1 Conv2d. 512, BN, LR
5 × 5 Conv2d. 512, BN, LR	4 × 1 ConvTrans2d. 512, BN, LR
5 × 5 Conv2d. 1024, BN, LR	4 × 1 ConvTrans2d. 256, BN, LR
AvgPool1d (3)	4 × 1 ConvTrans2d. 1, BN, LR
FC1 (2048)	squeeze a dimension 5 × 1 Conv1d.1
FC2 (512)	AdaptiveAvgPool1d (3200)
FC_NC (num_classes), FC_RF (2)	Output:(1,3200)
Output1:(num_classes), SoftMax	
Output2:(2), SoftMax (dim = -1)	

شکل ۱۱. معماری DCGAN ارائه شده در [۵۲] برای تشخیص گوینده



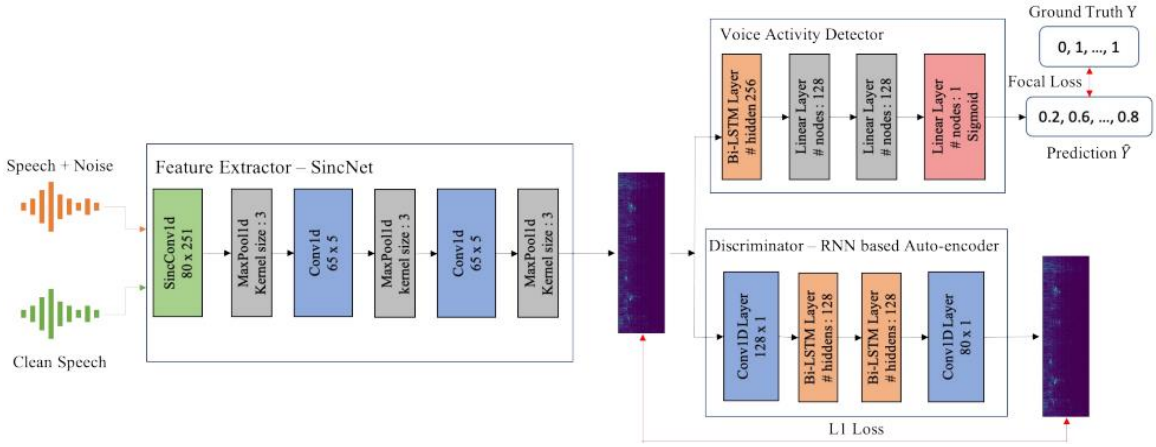
شکل ۱۲. معماری بکاررفته در [۵۳] برای کاربرد تشخیص فعالیت صوتی

مرجع [۵۳] برای کاربرد تشخیص فعالیت صوتی[۱۰۵] از فیلترهای سینک در یک معماری رقابتی استفاده کرده است. این مدل از لایه‌های بازگشتی LSTM برای تشخیص قطعات گفتاری سیگنال و دسته‌بندی دامنه استفاده می‌کند. شکل ۱۲ این معماری را نمایش می‌دهد.

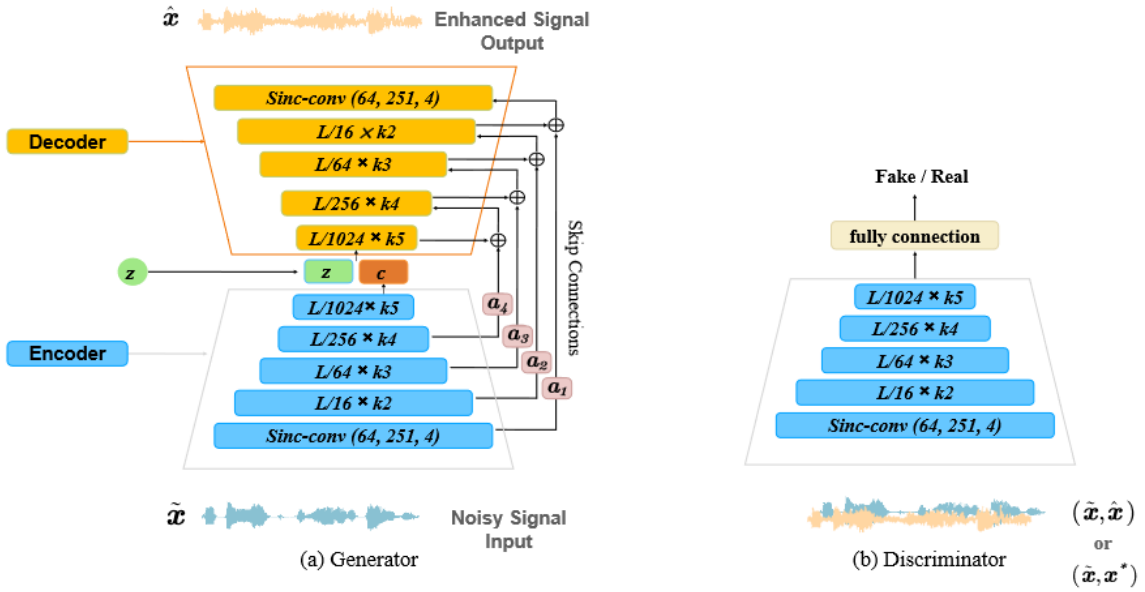
شبکه سیگنال خام ۲ ثانیه‌ای را به‌عنوان ورودی دریافت کرده و با استفاده از لایه‌های پیچشی سینک، ویژگی استخراج می‌کند. ویژگی‌های استخراج شده به دو شاخه تشخیص فعالیت صوتی و دسته‌بندی دامنه داده می‌شود. شاخه تشخیص فعالیت صوتی از یک پشته از LSTM های دوطرفه تشکیل شده است و بعد از آن سه لایه تمام متصل قرار دارند. شاخه دسته‌بندی دامنه از یک LSTM یک طرفه، تجمیع بیشینه[۱۰۶] در راستای محور زمان و یک شبکه تمام متصل تشکیل شده است.

در [۳۳] در هر دو قسمت مولد و جداساز یک معماری CGAN که برای تشخیص گوینده ارائه شده است، از لایه Sinc استفاده می‌کند. در اینجا نیز مشابه مرجع [۳۴] سه مجموعه فیلتر با طول‌های مختلف برای ساخت یک تصویر سه‌بعدی از سیگنال خام ورودی مورد استفاده قرار می‌گیرند. در ماژول مولد CGAN، بازنمایی حاصل به یک معماری رمزگذار-رمزگشا و در ماژول جداساز، به یک شبکه عصبی پیچشی داده می‌شود.

مرجع [۵۴] یک روش مبتنی بر معماری رقابتی مقاوم به نویز برای تشخیص فعالیت صوتی ارائه می‌دهد که ماژول جداساز آن یک خودرمزنگار مبتنی بر Bi-LSTM است. شکل ۱۳ این معماری را نمایش می‌دهد. در [۵۵] یک روش مبتنی بر GAN برای بهبود کیفیت گفتار ارائه شده است که در هر دو ماژول رمزکننده و رمزگشای ماژول مولد و نیز ماژول جداساز آن از لایه سینک استفاده شده است. این معماری که Sinc-SEGAN نام دارد، در شکل ۱۴ نمایش داده شده است.



شکل ۱۳. معماری رقابتی مقاوم به نویز ارائه شده در [۵۴] برای تشخیص فعالیت صوتی



شکل ۱۴. معماری Sinc–SEGAN ارائه شده در [۵۵] برای بهبود کیفیت گفتار

۴-۵. توابع زیان مورد استفاده

در برخی از معماری‌هایی که از فیلترهای پارامتری استفاده می‌کنند، از توابع زیان خاصی استفاده شده است که برخی از مهم‌ترین آن‌ها در این بخش معرفی می‌شوند.

تابع زیان بیشینه نرم[۱۰۷]، حاشیه زیادی بین دسته‌ها در نظر نگرفته و نمونه‌هایی که دسته‌بندی آن‌ها آسان یا مشکل است، را نادیده می‌گیرد. در [۵۶] برای حل این مشکلات از تکنیک مبتنی بر یادگیری کریکیلم[۱۰۸] استفاده شده است. در اینجا برای آموزش مدل سینک-نت از تابع زیان Curricular استفاده شده و معماری حاصل Curricular SincNet نام گرفته است. معادله ۲۴ این تابع زیان را نمایش می‌دهد. در این معادله، s پارامتر تغییر مقیاس و $\theta_{k,i}$ زاویه بین بردار وزن W_k و بردار ویژگی f_i است. پارامتر m، حاشیه اضافی است که در فضای زاویه‌ای فاصله بین دسته‌ها را افزایش می‌دهد.

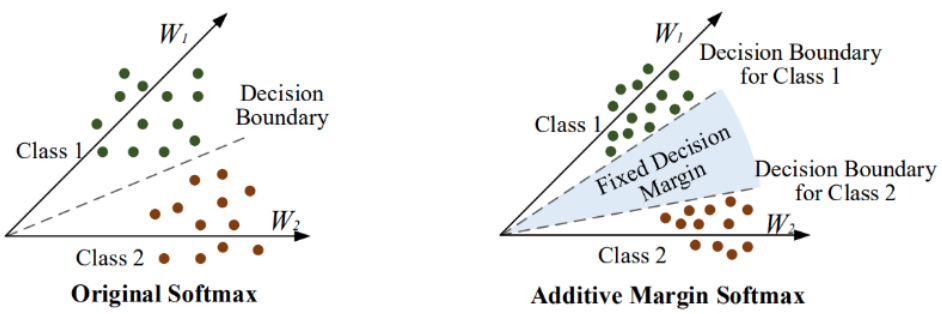
$$L_{\text{CurricularLoss}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s(\cos(\theta_{k,i}+m))}}{e^{s(\cos(\theta_{k,i}+m))} + \sum_{c=1,c\neq k}^{C} e^{sN(t,\cos\theta_c)}} \qquad (۲۴)$$

که در آن $N(t, \cos\theta_j)$ به‌صورت رابطه ۲۵ تعریف شده و برای ایجاد تمایز بین نمونه‌های آسان و مشکل در نظر گرفته می‌شود. در این معادله t ابرپارامتری است که با مقادیر نزدیک به صفر مقداردهی اولیه شده و در طی فرایند آموزش، مقدار آن افزایش می‌یابد.

$$N(t,\cos\theta_j) = \begin{cases} \cos\theta_j\,, \cos(\theta_k+m) > \cos\theta_j \\ \cos\theta_j(t+\cos\theta_j),\ \cos(\theta_k+m) < \cos\theta_j \end{cases} \qquad (۲۵)$$

مرجع [۴] برای افزایش کارایی مدل SincNet، تابع زیان بیشینه نرم با حاشیه اضافی[۱۰۹] (AM-Softmax) را معرفی می‌کند که از تابع زیان بیشینه نرم بهتر عمل می‌کند. این تابع، حاشیه جداسازی بین دسته‌ها در نظر می‌گیرد که باعث می‌شود نمونه‌های یک دسته به یکدیگر نزدیک‌تر و نمونه‌های دسته‌های متفاوت از یکدیگر دورتر قرار گیرند. این خاصیت باعث می‌شود، این تابع برای کاربردهای دسته‌بندی و تصدیق[۱۱۰] بهتر عمل کند. شکل ۱۵ این مفهوم را به‌صورت شماتیک نمایش می‌دهد. معادله AM-Softmax به‌صورت معادله ۲۶ است. در این معادله، $\boldsymbol{W}$ ماتریس وزن و f_i ورودی i اُمین نمونه از آخرین لایه تمام متصل است. s و m پارامترهایی هستند که به ترتیب مقدار مقیاس[۱۱۱] و حاشیه اضافی را تعیین می‌کنند.

$$L_{\text{AM-Softmax}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s(W_{y_i}^{\mathrm{T}} f_i - m))}}{\varphi_i + \sum_{j=1, j\neq y_i}^{c} e^{s(W_j^{\mathrm{T}} f_i))}} \qquad (۲۶)$$



شکل ۱۵. نمایش تفاوت تابع زیان بیشینه نرم و بیشینه نرم با حاشیه اضافی [۴]

۵. کاربردها

همان‌طور که در بخش‌های قبلی بیان شد، فیلترهای پارامتری در کاربردهای مختلف پردازش گفتار مورد استفاده قرار گرفته‌اند. در این بخش، مهم‌ترین مقالاتی که در آن‌ها از فیلترهای پارامتری استفاده شده، را از دیدگاه کاربردی تقسیم‌بندی کرده‌ایم. جدول ۳، لیست این مقالات را به تفکیک کاربرد به همراه سال ارائه، نوع فیلتر و معماری مورد استفاده نشان می‌دهد. همان‌طور که مشاهده می‌شود، استفاده از این نوع فیلترها در طیف وسیعی از کاربردهای پردازش گفتار مؤثر بوده و توانسته نسبت به لایه‌های پیچشی استاندارد به دقت‌های بهتری دست یابد.

جدول ۳. فیلترهای پارامتری و معماری‌های مورد استفاده در کاربردهای مختلف پردازش گفتار

کاربرد	نوع فیلتر	معماری	سال	مرجع
تشخیص فعالیت صوتی	Sinc	LSTM	۲۰۲۰	[۵۳]
	Gammachirp	convolutional attention network	۲۰۲۴	[۱۹]
	Sinc	CNN + LSTM	۲۰۲۲	[۴۷]
	Gabor	Efficient-LEAF	۲۰۲۳	[۳۶]
	Sinc	adversarial domain adaptive VAD	۲۰۲۲	[۵۴]
تشخیص ناهنجاری	Sinc	CNN	۲۰۲۳	[۳۷]
	Sinc	CNN	۲۰۲۲	[۳۸]
شناسایی گوینده	Sinc	CNN	۲۰۱۸	[۷]
	Sinc	RawNet	۲۰۲۰	[۳۰]
	Analytic	RawNet3	۲۰۲۲	[۲]
	Sinc	AM-CNN	۲۰۱۹	[۴]
	Complex Exponential	CVCNN	۲۰۲۱	[۳]
	Sinc	CNN	۲۰۲۳	[۵۷]
	Sinc	CNN	۲۰۲۲	[۵۸]
	IIRI	CNN	۲۰۲۳	[۲۶]
	Sinc, Gabor	ECAPA-TDNN	۲۰۲۲	[۴۱]
	Sinc	CNN	۲۰۲۱	[۵۹، ۶۰]
	Sinc	Curricular-CNN	۲۰۲۱	[۵۶]
	Sinc	CNN + x-vector	۲۰۲۰	[۶۱]
	Sinc	DCGAN	۲۰۲۲	[۵۲]
	Sinc	Sinc-Attention	۲۰۲۳	[۴۴]
	Sinc, $Sinc^2$, Gammatone, Gaussian, Cascade	CNN	۲۰۲۳	[۶]
	PF	CNN	۲۰۲۲	[۲۵]
شناسایی گفتار	Sinc	CNN	۲۰۱۸	[۷]

	Complex Gabor	CVCNN	۲۰۲۰	[۲۱]
	Sinc	E2E-SincNet	۲۰۲۰	[۴۶]
تشخیص تغییر گوینده	Sinc	LSTM	۲۰۲۲	[۴۸]
شمارش تعداد صداهای مجزا	Sinc	DyDecNet	۲۰۲۴	[۴۳]
تفکیک گوینده	Sinc	RNN	۲۰۲۰	[۶۲, ۶۳]
	Sinc	CNN	۲۰۲۱	[۶۴]
	Sinc	CNN	۲۰۱۹	[۶۵]
	Sinc	CNN	۲۰۲۳	[۶۶]
تشخیص احساس	Sinc	CNN	۲۰۲۰	[۶۷]
	Sinc	MS-SincResNet	۲۰۲۴	[۳۵]
	Sinc	CNN	۲۰۲۱	[۶۸]
بهبود کیفیت گفتار	Sinc	SDFCN	۲۰۲۰	[۲۷]
	Sinc	Sinc–SEGAN	۲۰۲۱	[۵۵]
	Gabor	WAV-UNET	۲۰۲۳	[۲۸]
	Sinc	CNN	۲۰۲۱	[۶۹]
تشخیص جنسیت گوینده	Sinc, $Sinc^2$, Gammatone, Gaussian, Cascade	CNN	۲۰۲۳	[۲۳]
	Sinc	CNN	۲۰۲۳	[۷۰]
ارزیابی کیفیت گفتار	Sinc	CNN-BLSTM	۲۰۲۴	[۷۱]
	Sinc, Gabor	CNN	۲۰۲۳	[۷۲]
	Sinc	CNN	۲۰۲۲	[۷۳]
Keyword Spotting	Sinc	CNN	۲۰۲۰	[۷۴]
	Sinc	CNN	۲۰۲۲	[۷۵]
	Sinc	CNN	۲۰۲۲	[۷۶]
	Sinc	CNN	۲۰۲۱	[۷۷]
تشخیص زبان گوینده	Sinc	CNN + RNN	۲۰۲۱	[۷۸]
	Sinc	CNN + RNN	۲۰۲۰	[۷۹]
تشخیص مکان گوینده	Sinc	CNN	۲۰۲۲	[۸۰]
	Sinc	SoundDet	۲۰۲۱	[۸۱]
موسیقی	Sinc	CNN	۲۰۲۱	[۸۲]
	Sinc	CNN + ResNet	۲۰۲۲	[۸۳]
	Sinc	MS-Sinc-ResNet	۲۰۲۱	[۳۴]
تشخیص بیماری	Sinc, $Sinc^2$, Gammatone, Gaussian, Parzen	CNN + LiGRU	۲۰۲۲	[۵۰]
	Sinc	CNN + LSTM + Attention	۲۰۲۰	[۵۱]
	Sinc	CNN	۲۰۲۳	[۸۴]
	Gabor	CNN + Attention	۲۰۲۳	[۴۵]
	Sinc	SINCNET-BIGRU	۲۰۲۳	[۸۵]
تقطیع	Sinc	CNN	۲۰۲۱	[۸۶]

جداسازی گفتار	analytic	CNN	۲۰۲۰	[۲۴]
	Gammatone, Gammachirp	CNN	۲۰۲۱	[۸۷]
دسته‌بندی صوت	Gabor	LEAF	۲۰۲۱	[۲۰]
	Gabor	LEAF	۲۰۲۳	[۳۶]
تشخیص گفتار جعلی	Sinc	RawNet2	۲۰۲۱	[۳۲]
	Sinc	RawGAT-ST	۲۰۲۱	[۸۸]
	Sinc	Differentiable Architecture Search	۲۰۲۱	[۸۹]
	Gabor	CNN	۲۰۲۳	[۹۰]

تاکنون مرجعی که عملکرد تمام فیلترهای معرفی شده در این مقاله را در یک معماری مشخص و با مجموعه داده یکسان مقایسه کند، ارائه نشده است اما برخی از مراجع مانند [۶, ۱۶, ۲۳, ۲۶] عملکرد زیرمجموعه‌ای از این فیلترها را در کاربردهایی مانند شناسایی گفتار، تشخیص گوینده و تعیین جنسیت گوینده مقایسه کرده‌اند. ازاین‌رو مقایسه جامع و عادلانه‌ای که بتوان بر اساس آن تصمیم گرفت کدام فیلتر برای چه کاربردی مناسب‌تر است، وجود ندارد. اما برای ایجاد یک نگرش کلی در مورد عملکرد این فیلترها، خلاصه‌ای از نتایجی که در این مراجع آمده است، توضیح داده می‌شود.

در مراجع [۶, ۲۶] اثر استفاده از فیلترهای استاندارد، میان‌گذر مستطیلی، مثلثی، گاماتون، گاوسی، آبشاری و IIR تغییریافته بر عملکرد یک شبکه عصبی پیچشی برای کاربرد تشخیص گوینده بررسی شده است. نتایج ارزیابی نشان می‌دهد که در این کاربرد، استفاده از فیلترهای آبشاری و IIR تغییریافته نتایج بهتر و فیلترهای استاندارد (فیلترهایی که ضرایب آن را خود شبکه یاد می‌گیرد) و میان‌گذر مستطیلی نتایج بدتری نسبت به سایر فیلترها داشته‌اند. در مرجع [۲۳] اثر استفاده از فیلترهای استاندارد، میان‌گذر مستطیلی، گاماتون، گاوسی و آبشاری در یک شبکه عصبی پیچشی برای کاربرد تشخیص جنسیت گوینده بررسی شده است. در این کاربرد، فیلتر گاماتون نسبت به سایر فیلترها به دقت بالاتری دست پیدا کرده است. استفاده از فیلترهای استاندارد، میان‌گذر مستطیلی، مثلثی، گاماتون و گاوسی برای کاربرد شناسایی واج در شبکه عصبی پیچشی در مرجع [۱۶] بررسی شده و نشان داده شده است که فیلتر مثلثی بهتر از بقیه فیلترها عمل کرده است. در همه این مراجع در مورد تفسیرپذیری فیلترهای پارامتری یادگرفته شده و حساسیت آن‌ها به فرکانس‌های مختلف بحث مفصلی انجام شده است.

مراجعی که به بررسی استفاده از بقیه فیلترها در کاربردهای دیگر پردازش گفتار پرداخته‌اند، عموماً عملکرد فیلتر را با فیلترهای استاندارد و میان‌گذر مستطیلی مقایسه کرده‌اند. به‌طورکلی، فیلترهای پارامتری به دلیل داشتن تعداد پارامترهای کمتر نسبت به فیلترهای استاندارد، بهتر عمل می‌کنند. اما در بین فیلترهای پارامتری، در اکثر کاربردها، فیلترهای میان‌گذر مستطیلی عملکرد بدتری نسبت به سایر فیلترهای پارامتری داشته‌اند.

۶. جمع‌بندی، نتیجه‌گیری و آینده نگری

روش‌های یادگیری ژرف توانسته‌اند به کارایی بالایی در

بسیاری از کاربردهای پردازش گفتار دست یابند. این افزایش کارایی در قبال کاهش تفسیرپذیری مدل حاصل شده است. به‌بیان‌دیگر، عملکرد داخلی معماری‌های ژرف مبهم بوده و ماهیت مدل جعبه سیاه است. در سال‌های اخیر توجه به جنبه تفسیرپذیری مدل‌های ژرف بیشتر شده است. تفسیرپذیری کمک می‌کند تا تصمیمات مدل تا حدی توجیه شده و بعلاوه، جنبه‌های مختلف مساله برای شخصی که برای حل آن تلاش می‌کند، شفاف‌تر شود. یادگیری فیلترهای معنی‌دار در لایه اول یک معماری ژرف، به تفسیرپذیری مدل کمک می‌کند. تفسیر فیلترهای یادگرفته شده در این لایه، نمایانگر حساسیت مدل نسبت به نواحی مختلف فرکانسی بوده و می‌تواند بازه‌های فرکانسی مهم را برای هر کاربرد تعیین کند. از آنجا که سامانه شنیداری انسان نیز قابلیت آنالیز فرکانسی دارد، مدل‌های فیلتری شنیداری مانند فیلترهای گاماتون نیز در این راه الهام‌بخش بوده است.

در این مقاله، انواع فیلترهای پارامتری که در معماری‌های ژرف برای کاربردهای مختلف پردازش گفتار مورد استفاده قرار گرفتند، معرفی شدند. در این بین فیلترهای میان‌گذر مستطیلی بیشترین استفاده را برده‌اند. ازآنجاکه این فیلترها در لایه اول معماری‌هایی که با ورودی سیگنال خام سروکار دارند، استفاده می‌شوند، پاسخ ضربه آن‌ها بررسی و ارائه شد.

پس‌ازآن انواع معماری‌هایی که در آن از این فیلترها استفاده شده است، معرفی شدند. در یک دسته‌بندی کلی می توان این معماری‌ها را در دسته‌های معماری‌های مبتنی بر شبکه عصبی پیچشی، مبتنی بر سازوکار توجه، مبتنی بر شبکه‌های عصبی بازگشتی و معماری‌های رقابتی جای داد. در وضعیت فعلی در اغلب کاربردها از شبکه عصبی پیچشی استفاده شده است.

در بخش پایانی مقاله، کاربردهایی از پردازش گفتار که از این فیلترها استفاده شده‌اند، معرفی شدند. این کاربردها دارای تنوع زیادی شامل تشخیص فعالیت صوتی، شناسایی گوینده، شناسایی گفتار، افزایش کیفیت گفتار، تفکیک گوینده، تشخیص جنسیت گوینده، تشخیص زبان گوینده، تشخیص مکان گوینده، تشخیص بیماری و تشخیص گفتار جعلی بودند. در بسیاری از این کاربردها، بهره‌گیری از فیلترهای پارامتری، علاوه بر افزایش کارایی مدل، با کم کردن تعداد پارامترهای آن به افزایش سرعت همگرایی و جلوگیری از بیش برازش کمک می‌کند.

ازآنجاکه استفاده از فیلترهای پارامتری در اکثر کاربردهایی که در اینجا بررسی شد، منجر به بهبود عملکرد مدل‌های یادگیری ژرف شده است، استفاده از آن‌ها در آینده گسترش بیشتری پیدا خواهد کرد. به‌خصوص استفاده از فیلترهای پارامتری در معماری‌های جدیدتر مانند مبدل‌ها، می‌تواند باعث بهبود عملکرد آن‌ها گردد. با وجود این که در بیشتر کاربردها، از فیلتر مستطیلی استفاده شده است اما عموماً، فیلتر گاماتون نسبت به سایر فیلترها، باعث بهبود بیشتر کارایی در مدل‌های ژرف شده است. ازاین‌رو، انتظار می‌رود، استفاده از فیلترهای خاص‌تر مانند گاماتون نسبت به فیلتر مستطیلی گسترش بیشتری پیدا کند.

از سوی دیگر، یادگیری فیلترهای معنی‌دار در لایه اول یک معماری ژرف تداعی‌کننده آنالیز فرکانسی گوش در کاربردهای مختلف پردازشی است. تجزیه‌وتحلیل بانک فیلتر یادگرفته شده در این لایه علاوه بر توضیح عملکرد خود مدل ژرف، می‌تواند عملکرد گوش انسان و حساسیت فرکانسی آن در کاربردهای مختلف پردازش گفتار را نمایان سازد.

۷. منابع

[۱] Lyon, Richard F. *Human and machine hearing: extracting meaning from sound*. Cambridge University Press, 2017.

[۲] Jung, Jee-weon, You Jin Kim, Hee-Soo Heo, Bong-Jin Lee, Youngki Kwon, and Joon Son Chung. "Pushing the limits of raw waveform speaker recognition." *arXiv preprint arXiv:2203.08488* (2022).

[۳] Peng, Junyi, Xiaoyang Qu, Jianzong Wang, Rongzhi Gu, Jing Xiao, Lukás Burget, and Jan Cernocký. "ICSpk: Interpretable Complex Speaker Embedding Extractor from Raw Waveform." In *Interspeech*, pp. 511-515. 2021.

[٤] Nunes, Joao Antônio Chagas, David Macêdo, and Cleber Zanchettin. "Additive margin sincnet for speaker recognition." In *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1-5. IEEE, 2019.

[٥] Fayyazi, Hossein, and Yasser Shekofteh. "IIRI-Net: An interpretable convolutional front-end inspired by IIR filters for speaker identification." *Neurocomputing* 558 (2023): 126767.

[٦] Fayyazi, Hossein, and Yasser Shekofteh. "Analyzing the Use of Auditory Filter Models for Making Interpretable Convolutional Neural Networks for Speaker Identification." In *2023 28th International Computer Conference, Computer Society of Iran (CSICC)*, pp. 1-6. IEEE, 2023.

[۷] Ravanelli, Mirco, and Yoshua Bengio. "Speaker recognition from raw waveform with sincnet." In *2018 IEEE spoken language technology workshop (SLT)*, pp. 1021-1028. IEEE, 2018.

[۸] Guidotti, Riccardo, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. "A survey of methods for explaining black box models." *ACM computing surveys (CSUR)* 51, no. 5 (2018): 1-42.

[۹] Ge, Wanying, Jose Patino, Massimiliano Todisco, and Nicholas Evans. "Explaining deep learning models for spoofing and deepfake detection with SHapley Additive exPlanations." In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6387-6391. IEEE, 2022.

[۱۰] Slack, Dylan, Anna Hilgard, Sameer Singh, and Himabindu Lakkaraju. "Reliable post hoc explanations: Modeling uncertainty in explainability." *Advances in neural information processing systems* 34 (2021): 9391-9404.

[۱۱] Agrawal, Purvi, and Sriram Ganapathy. "Interpretable representation learning for speech and audio signals based on relevance weighting." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020): 2823-2836.

[۱۲] Jiang, Junyan, Gus G. Xia, Dave B. Carlton, Chris N. Anderson, and Ryan H. Miyakawa. "Transformer vae: A hierarchical model for structure-aware and interpretable music representation learning." In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 516-520. IEEE, 2020.

[۱۳] Moore, Brian CJ. *An introduction to the psychology of hearing*. Brill, 2012.

[۱٤] Palaz, Dimitri, and Ronan Collobert. "Analysis of CNN-based speech recognition system using raw speech as input." (2015).

[۱٥] Rabiner, Lawrence, and Ronald Schafer. *Theory and applications of digital speech processing*. Prentice Hall Press, 2010.

[۱٦] Loweimi, Erfan, Peter Bell, and Steve Renals. "On learning interpretable CNNs with parametric modulated kernel-based filters." In *Interspeech 2019*, pp. 3480-3484. International Speech Communication Association, 2019.

[۱۷]Formby, C. "Simple triangular approximations of auditory filter shapes." *Journal of Speech, Language, and Hearing Research* 33, no. 3 (1990): 530-539.
[۱۸]Johannesma, P. L. M. "The pre-response stimulus ensemble of neurons in the cochlear nucleus." In *Symposium on Hearing Theory, 1972*. IPO, 1972.
[۱۹]Li, Nan, Longbiao Wang, Meng Ge, Masashi Unoki, Sheng Li, and Jianwu Dang. "Robust voice activity detection using an auditory-inspired masked modulation encoder based convolutional attention network." *Speech Communication* 157 (2024): 103024.
[۲۰]Zeghidour, Neil, Olivier Teboul, Félix De Chaumont Quitry, and Marco Tagliasacchi. "LEAF: A learnable frontend for audio classification." *arXiv preprint arXiv:2101.08596* (2021).
[۲۱]Noé, Paul-Gauthier, Titouan Parcollet, and Mohamed Morchid. "Cgcnn: Complex gabor convolutional neural network on raw speech." In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7724-7728. IEEE, 2020.
[۲۲]Oglic, Dino, Zoran Cvetkovic, Peter Bell, and Steve Renals. "A deep 2D convolutional network for waveform-based speech recognition." In *Interspeech 2020*, pp. 1654-1658. International Speech Communication Association, 2020.
[۲۳]Fayyazi, Hossein, and Yasser Shekofteh. "Exploiting auditory filter models as interpretable convolutional frontends to obtain optimal architectures for speaker gender recognition." *Applied Acoustics* 213 (2023): 109635.
Pariente, Manuel, Samuele Cornell, Antoine Deleforge, and Emmanuel Vincent. "Filterbank design for end-to-end speech separation." In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6364-6368. IEEE, 2020.
[۲۵]W. Li, Z. Tan, Z. Xia, D. Wu, and J. Ning, "PF-Net: Personalized Filter for Speaker Recognition from Raw Waveform," in *International Conference on Mobile Computing, Applications, and Services*, 2022: Springer, pp. 362-374 .
[۲۶]H. Fayyazi and Y. Shekofteh, "IIRI-Net: An interpretable convolutional front-end inspired by IIR filters for speaker identification," *Neurocomputing*, vol. 558, p. 126767, 2023.
[۲۷]C.-L. Liu, S.-W. Fu, Y.-J. Li, J.-W. Huang, H.-M. Wang, and Y. Tsao, "Multichannel speech enhancement by raw waveform-mapping using fully convolutional networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1888-1900, 2020.
[۲۸]F. Mathieu, T. Courtat, G. Richard, and G. Peeters, "Learning Interpretable Filters In Wav-UNet For Speech Enhancement," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023: IEEE, pp. 1-5 .
[۲۹]D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," *arXiv preprint arXiv:1806.03185*, 2018.
[۳۰]J.-w. Jung, S.-b. Kim, H.-j. Shim, J.-h. Kim, and H.-J. Yu, "Improved rawnet with feature map scaling for text-independent speaker verification using raw waveforms," *arXiv preprint arXiv:2004.00526*, 2020.
[۳۱]J.-w. Jung, H.-S. Heo, J.-h. Kim, H.-j. Shim, and H.-J. Yu, "Rawnet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification," *arXiv preprint arXiv:1904.08104*, 2019.
[۳۲]H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-end anti-spoofing with rawnet2," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021: IEEE, pp. 6369-6373 .
[۳۳]G. Wei, Y. Zhang, H. Min, and Y. Xu, "End-to-end speaker identification research based on multi-scale SincNet and CGAN," *Neural Computing and Applications*, vol. 35, no. 30, pp. 22209-22222, 2023.

[۳٤]P.-C. Chang, Y.-S. Chen, and C.-H. Lee, "MS-SincResnet: Joint learning of 1D and 2D kernels using multi-scale SincNet and ResNet for music genre classification," in *Proceedings of the 2021 International Conference on Multimedia Retrieval*, 2021, pp. 29-36 .
[۳٥]P.-C. Chang, Y.-S. Chen, and C.-H. Lee, "IIOF: Intra-and Inter-feature orthogonal fusion of local and global features for music emotion recognition," *Pattern Recognition*, vol. 148, p. 110200, 2024.
[۳٦]M. Anderson, T. Kinnunen, and N. Harte, "Learnable frontends that do not learn: Quantifying sensitivity to filterbank initialisation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023: IEEE, pp. 1-5 .
[۳۷]S. Kulkarni, H. Watanabe, and F. Homma, "Self-Supervised Audio Encoder with Contrastive Pretraining for Respiratory Anomaly Detection," in *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 2023: IEEE, pp. 1-5 .
[۳۸]D. Fedorishin *et al.*, "Large-Scale Acoustic Automobile Fault Detection: Diagnosing Engines Through Sound," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022, pp. 2871-2881 .
[۳۹]W. Ghezaiel, L. Brun, and O. Lézoray, "Hybrid network for end-to-end text-independent speaker identification," in *2020 25th International conference on pattern recognition (ICPR)*, 2021: IEEE, pp. 2352-2359 .
[٤۰]W. Ghezaiel, B. Luc, and O. Lézoray, "Wavelet scattering transform and CNN for closed set speaker identification," in *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*, 2020: IEEE, pp. 1-6 .
[٤۱]J. Li, Y. Tian, and T. Lee, "Learnable frequency filters for speech feature extraction in speaker verification," *arXiv preprint arXiv:2206.07563*, 2022.
[٤۲]B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," *Interspeech 2020*, 2020.
[٤۳]Y. He, Z. Dai, N. Trigoni, L. Chen, and A. Markham, "SoundCount: Sound Counting from Raw Audio with Dyadic Decomposition Neural Network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, vol. 38, no. 11, pp. 12421-12429 .
[٤٤]L. Li, J. Li, D. Wang, X. Wang, and S. Qiao, "Sinc-attention feature extraction for trivial-event based speaker verification," *Electronics Letters*, vol. 59, no. 9, p. e12812, 2023.
[٤٥]W. Yang *et al.*, "Attention guided learnable time-domain filterbanks for speech depression detection," *Neural Networks*, vol. 165, pp. 135-149, 2023.
[٤٦]T. Parcollet, M. Morchid, and G. Linares, "E2E-SINCNET: Toward fully end-to-end speech recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020: IEEE, pp. 7714-7718 .
[٤۷]Z. Du, K. Liu, X. Wan, and H. Zhou, "Joint speech activity and overlap detection with multi-exit architecture," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2022: IEEE, pp. 59-65 .
[٤۸]H. Su *et al.*, "A multitask learning framework for speaker change detection with content information from unsupervised speech decomposition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022: IEEE, pp. 8087-8091 .
[٤۹]K. Qian, Y. Zhang, S. Chang, M. Hasegawa-Johnson, and D. Cox, "Unsupervised speech decomposition via triple information bottleneck," in *International Conference on Machine Learning*, 2020: PMLR, pp. 7836-7846 .
[٥۰]Z. Yue, E. Loweimi, H. Christensen, J. Barker, and Z. Cvetkovic, "Dysarthric Speech Recognition From Raw Waveform with Parametric CNNs," in *INTERSPEECH*, 2022, pp. 31-35 .

[٥١]Y. Pan *et al.*, "Acoustic feature extraction with interpretable deep neural network for neurodegenerative related disorder classification," in *Proceedings of Interspeech 2020*, 2020: International Speech Communication Association (ISCA), pp. 4806-4810 .
[٥٢]Y. Zhang, G. Wei, H. Min, and Y. Xu, "Text-Independent Speaker Identification Using a Single-Scale SincNet-DCGAN Model," in *International Conference on Data Mining and Big Data*, 2022: Springer, pp. 18-28 .
[٥٣]M. Lavechin *et al.*, "End-to-end Domain-Adversarial Voice Activity Detection," in *Interspeech 2020*, 2020 .
[٥٤]T. Kim, J. Chang, and J. H. Ko, "Ada-vad: Unpaired adversarial domain adaptation for noise-robust voice activity detection," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022: IEEE, pp. 7327-7331 .
[٥٥]L. Li, Wudamu, L. Kuerzinger, T. Watzel, and G. Rigoll, "Lightweight End-to-End Speech Enhancement Generative Adversarial Network Using Sinc Convolutions," *Applied Sciences*, vol. 11, no. 16, p. 7564, 2021.
[٥٦]L. Chowdhury, M. Kamal, N. Hasan, and N. Mohammed, "Curricular sincnet: Towards robust deep speaker recognition by emphasizing hard samples in latent space," in *2021 International Conference of the Biometrics Special Interest Group (BIOSIG)*, 2021: IEEE, pp. 1-4 .
[٥٧]N. Shome, B. Saritha, R. Kashyap, and R. H. Laskar, "A robust DNN model for text-independent speaker identification using non-speaker embeddings in diverse data conditions," *Neural Computing and Applications*, vol. 35, no. 26, pp. 18933-18947, 2023.
[٥٨]B. Saritha, N. Shome, R. H. Laskar, and M. Choudhury, "Enhancement in speaker recognition using SincNet through optimal window and frame shift," in *2022 2nd International Conference on Intelligent Technologies (CONIT)*, 2022: IEEE, pp. 1-6 .
[٥٩]M. Guo, J. Yang, and S. Gao, "Speaker recognition method for short utterance," in *Journal of physics: conference series*, 2021, vol. 1827, no. 1: IOP Publishing, p. 012158 .
[٦٠]Z. Li and J. Whitehill, "Compositional embedding models for speaker identification and diarization with simultaneous speech from 2+ speakers," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021: IEEE, pp. 7163-71 .٦٧
[٦١]M. Tripathi, D. Singh, and S. Susan, "Speaker recognition using SincNet and X-vector fusion," in *Artificial Intelligence and Soft Computing: 19th International Conference, ICAISC 2020, Zakopane, Poland, October 12-14, 2020, Proceedings, Part I 1* :٢٠٢٠ ,٩Springer, pp. 252-260 .
[٦٢]H. Bredin *et al.*, "Pyannote. audio: neural building blocks for speaker diarization," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020: IEEE, pp. 7124-7128 .
[٦٣]L. Bullock, H. Bredin, and L. P. Garcia-Perera, "Overlap-aware diarization: Resegmentation using neural end-to-end overlapped speech detection," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 202 :٠IEEE, pp. 7114-7118 .
[٦٤]H. Bredin and A. Laurent, "End-to-end speaker segmentation for overlap-aware resegmentation," *arXiv preprint arXiv:2104.04045*, 2021.
[٦٥]H. Dubey, A. Sangwan, and J. H. Hansen, "Transfer learning using raw waveform sincnet for robust speaker diarization," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019: IEEE, pp. 6296-6300 .
[٦٦]J. M. Coria, H. Bredin, S. Ghannay, and S. Rosset, "Continual self-supervised domain adaptation for end-to-end speaker diarization," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, 2023: IEEE, pp. 626-632 .

[٦٧]D. Priyasad, T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Attention driven fusion for multi-modal emotion recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020: IEEE, pp. 3227-3231 .
[٦٨]A. Anand, S. Negi, and N. Narendra, "Filters Know How You Feel: Explaining Intermediate Speech Emotion Classification Representations," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2021: IEEE, pp. 756-761 .
[٦٩]Y.-J. Li, S.-S. Wang, Y. Tsao, and B. Su, "Mimo speech compression and enhancement based on convolutional denoising autoencoder," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2021: IEEE, pp. 1245-1250 .
[٧٠]K. Radha and M. Bansal, "Towards modeling raw speech in gender identification of children using sincNet over ERB scale," *International Journal of Speech Technology,* vol. 26, no. 3, pp. 651-663, 2023.
[٧١]R. E. Zezario, B.-R. B. Bai, C.-S. Fuh, H.-M. Wang, and Y. Tsao, "Multi-Task Pseudo-Label Learning for Non-Intrusive Speech Quality Assessment Model," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024: IEEE, pp. 831-835 .
[٧٢]C. O. Mawalim, B. A. Titalim, S. Okada, and M. Unoki, "Auditory Model Optimization with Wavegram-CNN and Acoustic Parameter Models for Nonintrusive Speech Intelligibility Prediction in Hearing Aids," in *2023 31st European Signal Processing Conference (EUSIPCO)*, 2023: IEEE, pp. 211-215 .
[٧٣]R. E. Zezario, S.-W. Fu, F. Chen, C.-S. Fuh, H.-M. Wang, and Y. Tsao, "Deep learning-based non-intrusive multi-objective speech assessment model with cross-domain features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 54-70, 2022.
[٧٤]S. Mittermaier, L. Kürzinger, B. Waschneck, and G. Rigoll, "Small-footprint keyword spotting on raw audio data with sinc-convolutions," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020: IEEE, pp. 7454-7458 .
[٧٥]D. Peter, W. Roth, and F. Pernkopf, "End-to-end keyword spotting using neural architecture search and quantization," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022: IEEE, pp. 3423-3427 .
[٧٦]D. Kim, K. Ko, D. K. Han ,and H. Ko, "Discriminatory and orthogonal feature learning for noise robust keyword spotting," *IEEE Signal Processing Letters,* vol. 29, pp. 1913-1917, 2022.
[٧٧]A. Mohanty, A. Frischknecht, C. Gerum, and O. Bringmann, "Behavior of keyword spotting networks under noisy conditions," in *International Conference on Artificial Neural Networks*, 2021: Springer, pp. 369-378 .
[٧٨]Y. Qian *et al.*, "Speech-language pre-training for end-to-end spoken language understanding," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021: IEEE, pp. 7458-7462 .
[٧٩]L. Yang, K. Fu, J. Zhang, and T. Shinozaki, "Pronunciation Erroneous Tendency Detection with Language Adversarial Represent Learning," in *INTERSPEECH*, 2020 ,pp. 3042-3046 .
[٨٠]A. Berg, M. O'Connor, K. Åström, and M. Oskarsson, "Extending gcc-phat using shift equivariant neural networks," *arXiv preprint arXiv:2208.04654,* 2022.
[٨١]Y. He, N. Trigoni, and A. Markham, "SoundDet: Polyphonic moving sound event detection and localization from raw waveform," in *International Conference on Machine Learning*, 2021: PMLR, pp. 4160-4170 .
[٨٢]H.-H. Wu *et al.*, "Multi-task self-supervised pre-training for music classification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021: IEEE, pp. 556-560 .

[۸۳]X. Shi, E. Cooper, and J. Yamagishi, "Use of speaker recognition approaches for learning and evaluating embedding representations of musical instrument sounds," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 30, pp. 367-377, 2022.
[۸٤]S. Sabesan, A. Fragner, C. Bench, F. Drakopoulos, and N. A. Lesica, "Large-scale electrophysiology and deep learning reveal distorted neural signal dynamics after hearing loss," *Elife,* vol. 12, p. e85108, 2023.
[۸٥]R. V. Sharan, K. Qian, and Y. Yamamoto, "Automated Cough Sound Analysis for Detecting Childhood Pneumonia," *IEEE Journal of Biomedical and Health Informatics,* 2023.
[۸٦]H. Bredin and A. Laurent, "End-to-end speaker segmentation for overlap-aware resegmentation," in *Interspeech 2021*, 2021 .
[۸۷]H. Li, K. Chen, and B. U. Seeber, "Auditory filterbanks benefit universal sound source separation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021: IEEE, pp. 181-185 .
[۸۸]H. Tak, J.-w. Jung, J. Patino, M. Kamble, M. Todisco, and N. Evans, "End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection," *arXiv preprint arXiv:2107.12710,* 2021.
[۸۹]W. Ge, J. Patino, M. Todisco, and N. Evans, "Raw differentiable architecture search for speech deepfake and spoofing detection," *arXiv preprint arXiv:2107.12212,* 2021.
[۹۰]B. Wickramasinghe, E .Ambikairajah, V. Sethu, J. Epps, H. Li, and T. Dang, "DNN controlled adaptive front-end for replay attack detection systems," *Speech Communication,* vol. 154, p. 102973, 2023.

پی‌نوشت

[1] SincNet
[2] Convolutional Neural Networks
[3] Attention mechanism
[4] Recurrent Neural Networks
[5] Generative Adversarial Networks
[6] transformers
[7] Black-box
[8] eXplainable Artificial Intelligence
[9] interpretability
[10] Post-hoc
[11] Transformer Variational Auto-Encoders
[12] SHapley Additive exPlanations
[13] Audi deepfake detection
[14] Multi-band
[15] Reissner membrane
[16] Basilar Membrane
[17] Characteristic Frequency
[18] Frequency Selectivity
[19] masking
[20] Fletcher
[21] Helmholtz
[22] Auditory filters
[23] Linear Time Invariant
[24] Rectangular bandpass filter
[25] Lowpass filter
[26] rectangular function
[27] phase distortions
[28] gain

[29] truncate
[30] ripple
[31] passband
[32] attenuation
[33] stopband
[34] windowing
[35] Hamming Window
[36] Hanning Window
[37] Blackman Window
[38] Kaiser
[39] Carrier
[40] Gammatone Filter
[41] revcor functions
[42] Gammachirp Filter
[43] Chirp
[44] Gabor Filter
[45] LEArnable Frontend
[46] resolution
[47] Parzen Window
[48] Epanechnikov
[49] Infinite impulse response
[50] truncate
[51] Forward-backward filtering
[52] Analytic filter
[53] envelope
[54] Short Time Fourier Transform
[55] shift-invariant representation
[56] parameterized analytic analysis filter
[57] Personalized Filter
[58] Deformation point
[59] Sharp cutoffs
[60] Frequency selectivity
[61] formants
[62] pitch
[63] smooting
[64] transient
[65] phonemes
[66] Fully Connected
[67] pooling
[68] Fully Convolutional Network
[69] dilated
[70] Gated Recurrent Unit
[71] Feature Map Scaling
[72] non-linear sigmoid function
[73] wideband
[74] Multi-scale filters
[75] stride
[76] Adaptive Average Pooling
[77] Self-supervised contrastive learning
[78] details
[79] approximation
[80] Auditory-inspired modulation encoder
[81] Auditory-inspired masked modulation encoder
[82] Convolutional attention network
[83] power envelope extraction
[84] modulation filterbank
[85] parallel convolutional neural network

[86] robust
[87] masking-weight estimator
[88] Learner block
[89] decoder
[90] Bidirectional Recurrent Neural Network
[91] encoder
[92] token
[93] Attention-based decoder
[94] symbol
[95] Long short-term memory
[96] multitask learning
[97] Unsupervised speech decomposition
[98] Content vector
[99] feed-forward
[100] Dysarthria
[101] Light Gated recurrent unit
[102] upsampling
[103] Transposed Convolution
[104] stride
[105] Voice Activity Detection
[106] max-pooling
[107] softmax
[108] Curriculum Learning
[109] Additive Margin Softmax
[110] verification
[111] scaling