دانشکدگان فنی دانشگاه تهران

یادگیری ماشین

گزارش اولیه

(بخش اول پروژه پایانی درس)

نام افراد گروه:

امید ملایی ۸۱۰۱۰۳۲۴۱

نادیه محمدی ۸۱۰۱۰۳۳۳۸

فاطمه صدری ۸۱۰۱۰۲۰۲۷

دی ۱۴۰۳

فهرست مطالب

۱ مقدمه‌ای بر voice authentication

۱_۱ اهمیت voice authentication

احراز هویت بیومتریک امروزه در صنایع مختلف کاربرد زیادی دارند و به داده‌های بیوگرافیکی مثل اثر انگشت، صدا، تصویر، اسکن عنبیه متکی هستند. احراز هویت صوتی یکی از انواع احراز هویت بیومتریک است که از گفتار کاربر برای تشخیص هویت و جنسیت او استفاده می‌کند. این فناوری با تحلیل الگوهای صوتی مانند زیر و بم، لحن، ریتم و فرکانس صحبت کردن، که برای هر فرد منحصربه‌فرد است، عمل می‌کند.

گفتار و صدای افراد می‌تواند بعنوان شاخصه‌ای منحصر به فرد برای آن‌ها باشد مثل اثرانگشت و اسکن چهره به همین دلیل احراز هویت صوتی مزایای بسیاری مانند ورود بدون تماس فیزیکی و مقاومت در برابر سرقت از طریق حملات مختلف نسبت به سایر روش‌های بیومتریک ارائه می‌دهد.

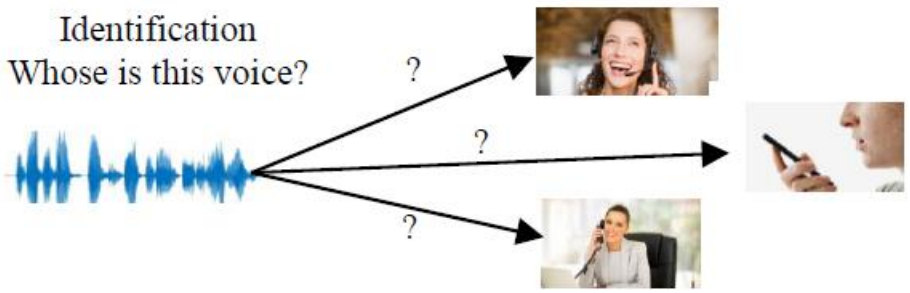
بزرگ‌ترین مزیت داده‌های بیومتریک این است که ایجاد نسخه‌های کپی شده و استفاده از آن‌ها برای احراز هویت بسیار دشوار است، چرا که این داده‌ها منحصربه‌فرد هستند. از نظر امنیتی، احراز هویت صوتی تأثیر بسیار بالاتری نسبت به تکنیک‌های سنتی احراز هویت و سایر ویژگی‌های بیومتریک دارد، زیرا مستقیماً با ورود به سیستم در تعامل است، نه فقط وارد کردن کد یا اطلاعات کاربری. از دیگر مزیت های احراز هویت صوتی می‌توان به دسترسی و احراز هویت سریع‌تر و آسان‌تر برای افراد با معلولیت‌های جسمی اشاره کرد.

احراز هویت صوتی نقش برجسته‌ای در حوزه‌های مختلف ایفا می‌کند از جمله کاربردهای آن می‌توان به شناسایی گوینده و تشخیص جنسیت گوینده اشاره کرد. این فناوری با استفاده از مدل‌های یادگیری ماشین و استخراج ویژگی‌های پیشرفته صوت، راه‌حل‌های امن و موثری ارائه می‌دهد.

۱_۲ کاربردهای voice authentication

۱_۲_۱ شناسایی گوینده(Speaker Identification)

شناسایی گوینده فرآیندی است که هویت یک شخص را با استفاده از صدای او تعیین می‌کند. ویژگی‌های صوتی هر فرد، مانند زیر و بم صدا، لحن و لهجه، به عنوان شاخص بیومتریک منحصربه‌فرد عمل می‌کنند. شناسایی گوینده در فرآیندهای امنیت و کنترل دسترسی(استفاده در سیستم‌های بانکی برای ورود به حساب کاربری یا تأیید تراکنش‌ها)، دستگاه‌های هوشمند و دستیارهای مجازی(Amazon Alexa, Google Assistant) و شناسایی مجرمان یا تأیید هویت مظنونین از صدای ضبط شده، از جمله کاربردهای احراز هویت صوتی است.



شکل ۱: شناسایی هویت با استفاده از صدای شخص

۱_۲_۲ تشخیص جنسیت گوینده(Gender Classification)

تشخیص جنسیت، تعیین جنسیت گوینده (مرد، زن، یا دیگر جنسیت‌ها) بر اساس ویژگی‌های صوتی به منظور تنظیم تعامل و پاسخ‌ها بر اساس جنسیت گوینده(سیستم‌های فعال‌شونده با صدا)، کمک به تشخیص مشکلات مرتبط با صدا یا شرایط درمانی(حوزه سلامت و مانیتورینگ)، تحلیل رفتار کاربران در مراکز تماس یا پلتفرم‌های رسانه‌ای(تحقیقات بازار) از جمله کاربردهای دیگر احراز هویت صوتی است.

از کاربردهای دیگر voice authentication:

- سازمان‌ها: قفل کردن و دسترسی به اطلاعات حیاتی و مکان‌های حساس.
- خانه‌های هوشمند: باز کردن قفل درها یا فعال کردن دستگاه‌ها تنها با صدای کاربر.

- تراکنش‌های امن: بسیاری از بانک‌ها از احراز هویت صوتی برای تأیید هویت مشتریان هنگام انجام تراکنش‌های مالی یا دریافت اطلاعات حساب استفاده می‌کنند.
- پشتیبانی مشتری: جایگزینی کلمات عبور برای دسترسی سریع و امن به اطلاعات مالی.
- پایش سلامت: تشخیص تغییرات در صدا که ممکن است نشان‌دهنده مشکلات سلامتی باشد.
- تشخیص زودهنگام: ابزارهای مبتنی بر صدا برای ارزیابی بیماری‌هایی نظیر پارکینسون یا افسردگی.
- فناوری هوشمند(IOT): احراز هویت برای دستگاه‌های متصل، مانند بلندگوهای هوشمند و سیستم‌های خودرو.
- آموزش مجازی و آموزش‌های آنلاین: تضمین هویت افراد شرکت‌کننده در آزمون‌ها برای جلوگیری از تقلب.

۱_۳ احراز هویت بسته و باز

احراز هویت محدود یا بسته(close-set authentication) تمام موجودیت‌ها (یا کلاس‌ها) که در فرایند احراز هویت استفاده می‌شوند، از پیش مشخص هستند. مدل برای یک مجموعه خاص از داده‌های شناخته شده آموزش دیده و اجرا می‌شود و داده‌های خارج از این محدوده را نمی‌پذیرد. کاربرد این روش در محیط‌های کنترل‌شده مثل بانک‌ها یا پایگاه‌های نظامی است.

احراز هویت باز(open-set authentication) در این رویکرد، سیستم برای مدیریت موجودیت‌های ناشناخته طراحی شده است. علاوه بر شناسایی داده‌های آشنا، ورودی‌های ناشناخته را نیز شناسایی کرده و رد می‌کند. این روش برای محیط‌های عمومی یا خدمات آنلاین مفید است که با تنوع بالای کاربران سروکار دارند. احراز هویت باز در سناریوهای پویاتری استفاده می‌شود که در آن سیستم برای مدیریت هم کاربران شناخته‌شده و هم ناشناخته طراحی شده است. در احراز هویت باز، سیستم می‌تواند بین کاربران مجاز (که در مجموعه شناخته‌شده هستند) و کاربران غیرمجاز (که خارج از مجموعه هستند) تمایز قائل شود. در صورتی

که یک فرد ناشناس بخواهد وارد سیستم شود، سیستم می‌تواند درخواست او را رد کرده یا هشدار دهد. این نوع احراز هویت معمولاً با دامنه وسیع‌تری از کاربران سروکار دارد.

۱_۳_۱ تفاوت های اصلی احراز هویت بسته و باز

تفاوت اصلی بین این دو روش در این است که سیستم‌های بسته فقط روی شناسایی سخنگوهای ثبت‌شده تمرکز دارند و از روش‌های طبقه‌بندی سنتی استفاده می‌کنند، در حالی که سیستم‌های باز باید بتوانند صدای افراد ناشناس را نیز به درستی شناسایی کرده و رد کنند. برای استفاده در کاربردهای دنیای واقعی مانند دستیارهای صوتی هوشمند یا سیستم‌های امنیتی، معمولاً از احراز هویت باز استفاده می‌شود، اما این کار به مدل‌های پیچیده‌تر نیاز دارد تا با مشکلات مرتبط با صدای ناشناسان کنار بیاید. به همین دلیل سیستم‌های احراز هویت باز می‌توانند سناریوهای متنوع‌تری را مدیریت کنند و قابلیت شناسایی افراد ناشناس یا غیرمجاز را دارند، در حالی که سیستم‌های محدود فقط روی کاربران از پیش تعریف‌شده تمرکز دارند. از نظر امنیتی نیز سیستم‌های محدود ممکن است امنیت بالاتری در محیط‌های کنترل‌شده ارائه دهند، اما سیستم‌های باز در کاربردهای وسیع‌تر و واقعی‌تر مفیدترند، جایی که تعامل با کاربران ناشناس رایج است.

۱_۳_۲ بررسی چگونگی پیاده سازی احراز هویت بسته

در این رویکرد، سیستم فقط سخنگوهایی را که در مجموعه‌ای از قبل مشخص، ثبت شده‌اند شناسایی می‌کند. سیستم نمونه‌ی صدای ورودی را با داده‌های کاربران ثبت‌شده مقایسه کرده و آن را به یکی از آن‌ها تعلق می‌دهد. یکی از روش‌های رایج در این سیستم‌ها استفاده از مدل‌های آی-وکتور (i-vector) یا شبکه‌های عصبی عمیق (DNN) است که ویژگی‌هایی مانند MFCC برای استخراج ویژگی‌های صدا به کار می‌روند. این سیستم از مدل‌های مدل مخلوط گاوسی (GMM) یا DNN برای اعتبارسنجی کاربران استفاده می‌کند.

برای پیاده‌سازی این روش مراحل کلیدی شامل موارد زیر است:

۱) پیش‌پردازش: استخراج ویژگی‌های صوتی مانند MFCC.

۲) ساخت مدل سخنگو: با استفاده از روش‌هایی مانند آی-وکتورها که نمایندگی‌های صوتی هر سخنگو را ایجاد می‌کند.

۳) آموزش مدل: یک طبقه‌بند برای شناسایی سخنگوهای ثبت‌شده آموزش داده می‌شود.

۱_۳_۳ بررسی چگونگی پیاده سازی احراز هویت باز

این روش به فراتر از شناسایی سخنگوهای شناخته‌شده می‌پردازد و باید همچنین توانایی شناسایی صدای افرادی که ثبت نشده‌اند و رد آن‌ها را نیز داشته باشد. این سیستم نه‌تنها برای شناسایی سخنگوهای مجاز بلکه برای تشخیص و رد سخنگوهای ناشناس طراحی شده است. یکی از چالش‌ها در این روش جلوگیری از شکست‌های شناسایی است که ممکن است صدای ناشناسی به اشتباه با مدل‌های ثبت‌شده تطابق پیدا کند. در سیستم‌های باز، معمولاً از روش‌های انطباق نمایه‌سازی (embedding adaptation) استفاده می‌شود، جایی که نمایه‌های سخنگو با داده‌های جدید انطباق داده می‌شود و از متدهایی مانند استفاده از تابع هزینه‌ی Contrastive loss برای بهینه‌سازی فاصله میان صدای ناشناس و شناخته‌شده بهره می‌برند.

مراحل پیاده‌سازی این روش شامل موارد زیر است:

۱) استخراج ویژگی‌ها: مشابه با روش قبل، از ویژگی‌هایی مانند MFCC یا نمایه‌های پیچیده‌تر برای ویژگی‌های صوتی استفاده می‌شود.

۲) انطباق نمایه‌سازی: این روش نمایه‌های سخنگو را انطباق داده و همزمان آن‌ها را از سخنگوهای ناشناس جدا می‌کند.

۳) اعتبارسنجی باز: هنگام اعتبارسنجی صدای سخنگو، سیستم باید به‌کمک معیارهایی مانند اندازه‌گیری انتروپی تصمیم بگیرد که آیا صدای دریافتی متعلق به سخنگوی شناخته‌شده است یا خیر.

۱_۳_۴ کاربردهای دو روش بالا در voice authentication

احراز هویت بسته بیشتر در محیط‌هایی استفاده می‌شود که فقط به افراد مشخص نیاز است تا به سرویس‌های امن دسترسی پیدا کنند، مانند برخی برنامه‌های مراقبت‌های بهداشتی یا خدمات مالی (مانند بانکداری تلفنی)، جایی که صحت شناسایی هویت اهمیت زیادی دارد.

احراز هویت باز معمولاً در شرایطی مانند خطوط خدمات مشتری خودکار یا دستگاه‌هایی مانند بلندگوهای هوشمند استفاده می‌شود که کاربران لزوماً قبلاً ثبت‌نام نکرده‌اند. این رویکرد امکان استفاده راحت‌تر برای گروه بزرگ‌تری از افراد را فراهم می‌کند اما نیاز به تدابیر اضافی برای مقابله با تقلب، مانند فناوری‌های ضد جعل، دارد تا از حملات استفاده از صداهای ضبط‌شده یا تولیدشده توسط ابزارهای شبیه‌سازی جلوگیری شود.

هر دو روش معمولاً با تشخیص زنده بودن تقویت می‌شوند تا اطمینان حاصل شود که شخصی که در حال احراز هویت است، خود واقعاً حضور دارد و ریسک حملات استفاده از صداهای ضبط‌شده یا سنتز شده را کاهش می‌دهد. بسته به کاربرد، کسب‌وکارها ممکن است انتخاب کنند که کدام یک از این روش‌ها را با توجه به تعادل میان انعطاف‌پذیری، امنیت و تعداد کاربران انتخاب کنند.

۲ بررسی چالش‌های voice authentication

۲_۱ چالش‌های احراز هویت صوتی

امنیت یکی از چالش‌های اساسی در احراز هویت صوتی است؛ زیرا ممکن است از صدای ضبط شده یا مصنوعی برای دور زدن امنیت استفاده شود. علاوه‌براین؛ عواملی مانند لهجه، وضعیت روحی، بیماری یا نویز محیطی می‌تواند روی عملکرد و قابلیت اطمینان احراز هویت صوتی تاثیر بگذارد.

اگر مجموعه داده‌هایی که برای آموزش سیستم‌های تشخیص هویت صوتی استفاده می‌شوند متنوع و گسترده نباشند، این سیستم‌ها نمی‌توانند با دقت بالا و به‌طور کلی کار کنند. به همین دلیل، داشتن داده‌های

متنوع برای آموزش این مدل‌ها که بتوانند در میان گروه‌های مختلف جمعیتی به خوبی عمل کنند بسیار مهم است.

۲_۲ چالش‌های طبقه‌بندی جنسیت بر اساس صوت

تشخیص جنسیت بر اساس صدا چالش‌های بسیاری دارد. برای مثال، بعضی صداها به دلیل ویژگی‌های مشترک مانند تن و زیروبم، گیج‌کننده هستند و تشخیص جنسیت را سخت می‌کنند. علاوه‌براین، اگر داده‌هایی که برای آموزش مدل‌های تشخیص جنسیت استفاده می‌شود متعادل نباشد و بیشتر از یک جنسیت باشد، مدل‌ها دچار بایاس می‌شوند و دقت طبقه‌بندی کاهش می‌یابد. در نهایت، تغییرات فرهنگی و زبانی مانند لهجه‌ها و الگوهای گفتاری مختلف در فرهنگ‌ها، ساخت یک سیستم تشخیص هویت صوتی جامع و دقیق را پیچیده‌تر می‌کند.

۲_۳ بررسی راه‌حل‌های بالقوه و تحقیقات جاری برای غلبه بر این چالش‌ها

راه‌حل‌های بالقوه و تحقیقات جاری برای غلبه بر چالش‌های احراز هویت صوتی و طبقه‌بندی جنسیتی شامل پیشرفت در الگوریتم‌های یادگیری ماشین، بهبود داده‌ها، و افزایش سازگاری مدل‌ها با ویژگی‌های صدای متغیر هستند که در ادامه به بررسی می‌پردازیم.

۲_۳_۱ راه‌حل غلبه بر چالش‌های احراز هویت بیومتریک صدا

یکی از چالش‌های عمده، تعصب موجود در داده‌های آموزشی است. بسیاری از سیستم‌های شناسایی صدای موجود، گروه‌هایی مانند افراد ترنس‌جندر و متنوع جنسیتی را کمتر نمایندگی می‌کنند که منجر به نتایج نادرست برای این گروه‌ها می‌شود، به ویژه برای افرادی که صدای آن‌ها به دلیل درمان‌های پزشکی یا عوامل دیگر تغییر کرده است. برای حل این مشکل، برخی از محققان به بهبود مجموعه داده‌ها برای نمایندگی بهتر این گروه‌ها و توسعه الگوریتم‌های تطبیقی که بتوانند ویژگی‌های صدای پویا را در نظر بگیرند، توصیه می‌کنند.

۲_۳_۲ راه‌حل غلبه بر مشکلات طبقه‌بندی جنسیتی

طبقه‌بندی جنسیتی تنها بر اساس صدا نیز با چالش‌هایی مواجه است. ویژگی‌های آکوستیک که به طور سنتی تفاوت‌های صدای مردانه و زنانه را مشخص می‌کنند(مانند فرکانس، نوا و بلندی) می‌توانند تحت تأثیر عوامل مختلف قرار گیرند و باعث دشواری در طبقه‌بندی دقیق شوند. علاوه بر این، داده‌های صوتی در زبان‌ها، لهجه‌ها و محیط‌های مختلف می‌توانند منجر به ناهماهنگی‌هایی شوند. راه‌حل پیشنهادی استفاده از یادگیری تجمیعی و الگوریتم‌های یادگیری نیمه‌نظارتی است که مدل‌های مختلف را برای بهبود دقت و پایداری ترکیب می‌کنند و به کمبود داده‌ها پاسخ داده و کارایی طبقه‌بندی را به طور کلی افزایش می‌دهند.

۲_۳_۳ مدیریت تغیرات صدا

چالش دیگر، پویایی طبیعت صدای گوینده است. عواملی مانند بیماری، استرس یا پیری می‌توانند ویژگی‌های صدا را به طور قابل توجهی تغییر دهند و منجر به شکست در احراز هویت یا طبقه‌بندی اشتباه شوند. برخی تحقیقات بر روی طراحی سیستم‌هایی تمرکز دارند که به مرور زمان یاد می‌گیرند و به این تغییرات تطبیق می‌یابند و عملکرد خود را با در نظر گرفتن ویژگی‌های بلندمدت صدا بهبود می‌بخشند.

۳ پیش‌پردازش داده‌های صوتی و اهمیت آن

صوت یکی از مهم‌ترین شکل‌های داده در طبیعت است که اطلاعات غنی و متنوعی را در خود جای داده است. استخراج ویژگی‌های صوتی، فرآیندی است که سیگنال صوتی را به مجموعه‌ای از مقادیر عددی قابل استفاده در مدل‌های یادگیری ماشین یا تحلیل سیستماتیک تبدیل می‌کند. این ویژگی‌ها نقش کلیدی در سیستم‌های پردازش صوت ایفا می‌کنند و کاربردهای گسترده‌ای در زمینه‌هایی چون تشخیص هویت صوتی، طبقه‌بندی جنسیت، تحلیل احساسات صوتی و تشخیص ژانر موسیقی دارند. در این میان، استفاده از روش‌های مناسب برای بازنمایی صوت و استخراج ویژگی، دقت مدل‌ها و عملکرد سیستم‌ها را به طرز چشمگیری افزایش می‌دهد.در این بخش به بررسی ساختارهای کلیدی برای بازنمایی صوت، پیش‌پردازش داده‌های صوتی، و

استخراج ویژگی‌های خاصی چون Log Mel Spectrogram، MFCC و Spectral Centroid می پردازیم. این روش‌ها به‌طور خاص برای بهبود کارایی سیستم‌های پردازش صوتی طراحی شده‌اند و در سیستم‌های تشخیص صدا و طبقه‌بندی جنسیت، به دلیل دقت بالا و تطابق مناسب با ویژگی‌های شنیداری انسان، اهمیت بسیاری دارند. برای مثال، استفاده از بازنمایی‌های فرکانسی مانند اسپکتروگرام و ویژگی‌های مشتق‌شده‌ای نظیر ضرایب سپسترال فرکانس مل (MFCC) ، امکان تحلیل دقیق‌تر و استخراج اطلاعات مرتبط‌تر را فراهم می‌کند.

بازنمایی صوت گامی کلیدی در سیستم‌های پردازش سیگنال محسوب می‌شود و هدف آن، تبدیل سیگنال خام به قالبی است که امکان استخراج ویژگی‌های معنادار را فراهم کند. در بسیاری از کاربردهای صوتی، از جمله تشخیص گفتار، طبقه‌بندی رویدادهای صوتی و بازشناسی گوینده، سیگنال خام که در حوزه زمان نمونه‌برداری می‌شود، مستعد نویزها و تغییرات محیطی است و همین موضوع تحلیل مستقیم آن را دشوار می‌سازد. از این رو، تلاش‌های متعددی صورت گرفته تا با توسعه روش‌های بازنمایی و استخراج ویژگی، اطلاعات ارزشمند نهفته در سیگنال به‌شکلی فشرده و درعین‌حال گویا نمایش داده شود. در این میان، استفاده از روش‌هایی مانند اسپکتروگرام، Log Mel Spectrogram، MFCC و ویژگی‌های طیفی، با هدف بهبود کارایی مدل‌های یادگیری ماشین بسیار رایج شده است.

اسپکتروگرام یکی از روش‌های اولیه اما همچنان پرکاربرد برای نمایش توزیع انرژی سیگنال در حوزه زمان-فرکانس است. در این روش، تبدیل فوریه کوتاه‌مدت (STFT) بخش‌های کوتاهی از سیگنال را به حوزه فرکانس می‌برد و بدین‌وسیله انرژی هر بازه زمانی در فرکانس‌های مختلف به‌صورت دیداری قابل‌تحلیل می‌شود. مزیت اصلی اسپکتروگرام درک الگوهای تکراری نظیر هارمونیک‌ها، تغییرات شدت فرکانس‌ها و ریتم است. با وجود این، تفکیک فرکانس‌های پایین در مقابل فرکانس‌های بالا در اسپکتروگرام معمول ممکن است دقیق نباشد؛ از همین رو استفاده از مقیاس مل (Log Mel Spectrogram) مطرح می‌شود تا مطابق با درک انسان از صدا، فرکانس‌های پایین با جزئیات بیشتری نمایش داده شوند. مقیاس مل، محدوده فرکانس را به‌گونه‌ای تقسیم می‌کند که تفاوت‌های فرکانس پایین برجسته‌تر و تفاوت‌های فرکانس بالا فشرده‌تر نمایش داده شوند.

در نتیجه، Log Mel Spectrogram به‌ویژه در کاربردهایی مانند شناسایی گوینده یا طبقه‌بندی عواطف صوتی بر پایه گفتار، دقت بالاتری نسبت به اسپکتروگرام خام ارائه می‌دهد.

علاوه بر Log Mel Spectrogram، روش MFCC (Mel Frequency Cepstral Coefficients) نیز در حوزه تشخیص گفتار و بازشناسی گوینده بسیار محبوب است. این روش ابتدا با بهره‌گیری از تبدیل فوریه کوتاه‌مدت، سیگنال را به دامنه فرکانس برده و سپس از بانک فیلترهای مل برای کاهش دامنه و نزدیک‌کردن بازه فرکانس به مقیاس گوش انسان استفاده می‌کند. آن‌گاه با اعمال تبدیل معکوس کسینوسی (DCT)، ضرایب کپسترالِ فشرده‌ای به دست می‌آید که معمولاً با حذف ضرایب کپسترال بالایی، اثر نویزهای محیطی کاهش می‌یابد. نتیجه نهایی، مجموعه‌ای کم‌حجم اما غنی از اطلاعات آکوستیکی است که برای الگوریتم‌های یادگیری ماشین مناسب‌اند. در مطالعات اخیر، ادغام MFCC با فیلترهای گابور نیز بررسی شده است تا اطلاعات دقیق‌تری از تغییرات زمانی-فرکانسی کسب شود. این رویکرد نشان می‌دهد که ترکیب چندین روش می‌تواند به استحکام بیشتری در برابر تنوع سیگنال‌ها و شرایط رکورد منجر شود.

از سوی دیگر، برخی ویژگی‌های طیفی مانند Spectral Centroid یا مرکز ثقل طیف، نقش مهمی در تحلیل تمبر یا رنگ صوتی ایفا می‌کنند. این ویژگی با سنجش محل تمرکز انرژی در حوزه فرکانس، معیاری برای «شفافیت» یا «تیرگی» صدا ارائه می‌دهد. در موسیقی، جنس صدا یا روشنایی صدا یکی از ابعاد بنیادین در تفکیک سازها و حتی سبک‌های موسیقی متفاوت است. ازاین‌رو، ترکیب Spectral Centroid با سایر ویژگی‌های آکوستیکی می‌تواند به مدلی جامع‌تر در طبقه‌بندی موسیقی یا تشخیص عواطف صوتی بینجامد.

در سال‌های اخیر، استفاده از روش‌های یادگیری عمیق همراه با بازنمایی‌های متنوع صوت گسترش یافته است. شبکه‌های عصبی پیچشی (CNN) با ورودی اسپکتروگرام یا Log Mel Spectrogram، قادر به یادگیری خودکار الگوهای زمانی-فرکانسی هستند و به‌ویژه در کاربردهایی مانند طبقه‌بندی ژانر موسیقی یا تحلیل کیفیت گفتار موفق عمل می‌کنند. برای نمونه، در پژوهشی از Convolutional Restricted Boltzmann Machine به‌عنوان روش استخراج ویژگی پیش‌پردازش شده استفاده شده و نتایج نشان داده است که این ترکیب به شناسایی مناسب‌تر ژانر موسیقی منجر می‌شود. همچنین، تلفیق روش‌های کلاسیک نظیر MFCC

با شبکه‌های عمیق یا مدل‌های مبتنی بر توالی و حافظه نظیر LSTM و GRU می‌تواند از الگوهای طولانی‌مدت در سیگنال سخنرانی بهره بگیرد و نرخ بازشناسی گوینده یا محتوای گفتار را افزایش دهد.

از منظر مهندسی کاربردی، فشرده‌سازی و ساده‌سازی ویژگی‌ها، علاوه بر دقت، نقشی اساسی در سرعت پردازش دارند. در کاربردهایی نظیر سامانه‌های بلادرنگ (Real-Time) یا دستگاه‌های پوشیدنی، محدودیت منابع محاسباتی و توان الکتریکی وجود دارد و در نتیجه، استفاده از مجموعه ویژگی‌های کوچک‌تر که درعین‌حال محتوای معنادار را حفظ کند، ضرورت می‌یابد. همچنین در محیط‌های پرنویز، افزایش مقاومت روش استخراج ویژگی در برابر نویز می‌تواند کیفیت عملکرد سیستم را ارتقا بخشد. روش‌هایی نظیر چندتابِر (Multitaper) در تبدیل فوریه یا به‌کارگیری تکنیک‌های تقویت داده از جمله تکرارهای کوتاه و تغییرات طیفی مصنوعی، راهکارهای بالقوه‌ای برای حل این مشکل هستند.

در نهایت، انتخاب بازنمایی مناسب نقش تعیین‌کننده‌ای در موفقیت سیستم‌های پردازش صوت دارد . هر یک از روش‌ها، اعم از اسپکتروگرام، Log Mel Spectrogram و MFCC و یا ویژگی‌های طیفی مانند Spectral Centroid ، مزایا و معایب خاص خود را دارد. در کاربردی مانند تشخیص گفتار پیوسته، روش‌هایی که بر اساس مقیاس مل ضرایب کپسترال را استخراج می‌کنند، سال‌هاست استانداردهای صنعتی را شکل داده‌اند. از سوی دیگر، در بازشناسی سازهای موسیقی، ترکیب ویژگی‌های تمبروال و طیفی با روش‌های یادگیری عمیق کارآمدتر به نظر می‌رسد. در مجموع، گسترش روزافزون تحقیقات نشان می‌دهد که تمرکز اصلی بر ابداع روش‌های انعطاف‌پذیر و چندلایه در بازنمایی صوت است؛ بدین ترتیب می‌توان با ترکیب ویژگی‌های کلاسیک و تکنیک‌های مبتنی بر شبکه‌های عمیق، سیستم‌هایی ساخت که ضمن حفظ سرعت و دقت بالا، در برابر نویز و تغییرات گسترده محیطی نیز مقاوم باشند. این رویکرد همه‌جانبه بی‌شک مسیر تحقیقات آینده را در حوزه پردازش صوت تعیین خواهد کرد و موجب شکل‌گیری نسل جدیدی از سیستم‌های تشخیص گفتار، طبقه‌بندی صوتی، و بازشناسی پیچیده‌تر ویژگی‌های آکوستیکی خواهد شد.

در سال‌های اخیر، پیشرفت‌های چشمگیری در حوزه پردازش صوت و گفتار صورت گرفته است و سیستم‌های مبتنی بر یادگیری ماشین و شبکه‌های عصبی عمیق توانسته‌اند عملکرد بسیار خوبی در وظایفی

نظیر تشخیص گفتار، بازشناسی گوینده و تحلیل احساسات صوتی از خود نشان دهند. با این حال، موفقیت این مدل‌ها تا حد زیادی به کیفیت داده‌های ورودی و به‌ویژه مرحله پیش‌پردازش سیگنال‌های صوتی وابسته است . پیش‌پردازش، شامل مجموعه‌ای از روش‌ها برای بهبود خلوص سیگنال و حذف عناصر نامربوط یا نویزی است که می‌تواند تأثیر چشمگیری در استخراج ویژگی‌های دقیق‌تر داشته باشد. در حقیقت، سیگنال‌های خام ضبط‌شده از محیط‌های واقعی غالباً حاوی نویز، اعوجاج فرکانسی و نوسانات دامنه هستند که اگر بدون اصلاحات لازم به الگوریتم‌های یادگیری ماشین ارائه شوند، موجب افت دقت یا حتی ناکارآمدی این سامانه‌ها می‌گردند. از این رو، شناخت و به‌کارگیری روش‌های مؤثر پیش‌پردازش گامی ضروری برای اطمینان از عملکرد مناسب سیستم‌های پردازش صوت است.

۳_۱ کاهش نویز / Noise Reduction

کاهش نویز (Noise Reduction) یکی از مراحل حیاتی پیش‌پردازش داده‌های صوتی است که با هدف حذف مؤثر صداهای مزاحم و افزایش خلوص سیگنال انجام می‌شود. نویزهای محیطی، مانند صدای پس‌زمینه، همهمه، و اعوجاج‌های فرکانسی، می‌توانند اطلاعات ارزشمند موجود در سیگنال صوتی را تضعیف کرده و دقت الگوریتم‌های پردازش صوت را کاهش دهند. حذف نویز، نه تنها کیفیت سیگنال را بهبود می‌بخشد، بلکه استخراج ویژگی‌هایی نظیر MFCC و Spectral Centroid را نیز دقیق‌تر می‌کند. به‌ویژه در کاربردهایی مانند تشخیص گوینده و طبقه‌بندی جنسیت، که هر نوع تغییر در سیگنال می‌تواند به دقت مدل آسیب بزند، کاهش نویز اهمیت فوق‌العاده‌ای دارد.

یکی از روش‌های متداول کاهش نویز، استفاده از فیلترهای باندی است که به حذف بخش‌هایی از سیگنال صوتی که خارج از محدوده فرکانسی مورد نظر هستند، می‌پردازد. این تکنیک، به‌ویژه برای حذف نویزهایی که در فرکانس‌های مشخصی مانند صدای هیس یا زمزمه پدیدار می‌شوند، بسیار مؤثر است. علاوه بر این، استفاده از روش‌های مبتنی بر کاهش نویز موجک (Wavelet Denoising) نیز بسیار رایج است. در این روش، سیگنال صوتی به چندین مقیاس زمانی-فرکانسی تجزیه می‌شود و نویز با آستانه‌گذاری نرم و سخت در

مقیاس‌های مختلف حذف می‌شود، در حالی که ساختار اصلی سیگنال حفظ می‌گردد. این تکنیک به دلیل کارایی بالا در حفظ ویژگی‌های مهم صوتی، به‌ویژه در داده‌های پرنویز محیطی، کاربرد زیادی دارد.

یکی دیگر از تکنیک‌های پیشرفته در کاهش نویز، استفاده از فیلترهای گابور چندمقیاسی است. این فیلترها با توانایی بالا در تحلیل سیگنال‌های صوتی در مقیاس‌های مختلف زمانی و فرکانسی، به‌طور همزمان به تقویت اجزای اصلی سیگنال و کاهش نویزهای ناخواسته کمک می‌کنند. به‌عنوان نمونه، ترکیب فیلتر گابور با روش کاهش نویز موجک در پروژه‌های پیچیده‌ای نظیر تشخیص گفتار و تشخیص گوینده، توانسته است دقت مدل‌ها را به طور قابل‌توجهی افزایش دهد.

روش‌های تطبیقی نیز نقش کلیدی در کاهش نویز دارند. در این رویکردها، الگوریتم‌ها با یادگیری ویژگی‌های نویز در محیط، به‌صورت پویا فیلترهایی را تنظیم می‌کنند که اجزای ناخواسته را حذف کنند. این تکنیک در محیط‌هایی که نویز غیرقابل پیش‌بینی و متغیر است، نظیر مکان‌های عمومی یا تماس‌های تلفنی، بسیار موثر است. همچنین، رویکردهای مبتنی بر یادگیری عمیق نظیر استفاده از شبکه‌های عصبی برای کاهش نویز، در سال‌های اخیر رشد چشمگیری داشته‌اند. این شبکه‌ها با آموزش روی مجموعه داده‌های نویزی، توانایی تفکیک نویز از سیگنال را با دقت بسیار بالا پیدا می‌کنند. کاهش نویز نه تنها سیگنال‌های تمیزتر و قابل‌اعتمادتری برای پردازش فراهم می‌کند، بلکه پایه‌ای قوی برای استخراج ویژگی‌هایی فراهم می‌سازد که در تحلیل‌های پیشرفته‌تر، نقش تعیین‌کننده‌ای دارند. بهبود دقت در تحلیل‌های صوتی، کاهش خطاهای تشخیص، و افزایش کارایی مدل‌های یادگیری ماشین از نتایج مستقیم این مرحله مهم هستند.

۳_۲ نرمال سازی / Normalization

نرمال‌سازی یکی از گام‌های کلیدی در پیش‌پردازش داده‌های صوتی است که هدف آن تعدیل دامنه سیگنال صوتی برای قرار گرفتن در یک محدوده ثابت است. این فرآیند به همگن‌سازی داده‌های ورودی کمک می‌کند و تأثیرات تفاوت‌های دامنه‌ای میان سیگنال‌ها را که ممکن است در مراحل پردازش و تحلیل اختلال ایجاد کند، به حداقل می‌رساند. در سیستم‌های یادگیری ماشین، نرمال‌سازی برای جلوگیری از تحت تأثیر

قرار گرفتن مدل توسط سیگنال‌های با دامنه بالا یا پایین ضروری است. به‌طور خاص، در کاربردهایی مانند تشخیص جنسیت و احراز هویت گوینده، نرمال‌سازی تضمین می‌کند که اطلاعات کلیدی صوتی، نظیر فرکانس‌های اصلی، بدون تحریف دامنه‌ای پردازش شوند. این مرحله معمولاً با استفاده از روش‌هایی مانند مقیاس‌بندی خطی انجام می‌شود که دامنه سیگنال را به محدوده مشخصی، نظیر [−1,1][1,1-][1,1−] یا [0,1][0,1][0,1]، نگاشت می‌کند. این روش تضمین می‌کند که تمامی سیگنال‌ها در یک مقیاس یکسان پردازش شوند و هیچ بخشی از داده به دلیل بزرگی یا کوچکی دامنه نادیده گرفته نشود. نرمال‌سازی علاوه بر ایجاد انسجام در داده‌ها، حساسیت مدل‌های یادگیری ماشین به تغییرات دامنه را کاهش می‌دهد، که به‌ویژه در وظایف مرتبط با طبقه‌بندی جنسیت و تشخیص گوینده بسیار حیاتی است. علاوه بر اسکالینگ خطی، روش‌های پیشرفته‌تری مانند نرمال‌سازی میانگین صفر و واریانس واحد (zzz-نرمال‌سازی) نیز به کار گرفته می‌شوند. در این روش، میانگین سیگنال از هر نمونه کم می‌شود و مقدار حاصل بر انحراف معیار تقسیم می‌شود. این فرآیند سیگنال را به گونه‌ای تعدیل می‌کند که دارای میانگین صفر و واریانس واحد باشد، که در مدل‌های حساس به تغییرات مقیاس، نظیر شبکه‌های عصبی، تأثیر مثبت قابل‌توجهی دارد. این روش به‌ویژه در داده‌های صوتی که تفاوت‌های گسترده‌ای در شدت صدا میان نمونه‌ها وجود دارد، مفید است. نرمال‌سازی همچنین با هموارسازی دامنه سیگنال، امکان استخراج ویژگی‌هایی نظیر ضرایب MFCC و Spectral Centroid را بهبود می‌بخشد. این ویژگی‌ها وابسته به دامنه نیستند و برای تمرکز بر جنبه‌های زمانی و فرکانسی سیگنال طراحی شده‌اند. به همین دلیل، نرمال‌سازی دامنه، پیش‌نیازی ضروری برای استخراج دقیق این ویژگی‌ها محسوب می‌شود. در کاربردهایی مانند تشخیص گفتار و گوینده، نرمال‌سازی نه تنها باعث بهبود عملکرد سیستم‌های پردازش صوتی می‌شود، بلکه اطمینان می‌دهد که مدل‌های یادگیری ماشین از داده‌هایی استفاده می‌کنند که به‌طور بهینه برای تحلیل آماده شده‌اند. این فرآیند همچنین به کاهش وابستگی نتایج به شرایط ضبط صوت، مانند فاصله میکروفون یا شدت صدای گوینده، کمک شایانی می‌کند. بنابراین، نرمال‌سازی نه تنها مرحله‌ای اساسی در پیش‌پردازش، بلکه شرطی حیاتی برای عملکرد مؤثر مراحل بعدی تحلیل صوت است.

۳_۳ پنجره بندی (windowing)

پنجره‌بندی یکی از مراحل اساسی در پیش‌پردازش سیگنال‌های صوتی است که به‌ویژه در تحلیل‌های زمانی-فرکانسی نقش کلیدی دارد. این فرآیند شامل تقسیم سیگنال صوتی به بخش‌های کوچک‌تر، یا پنجره‌ها، است تا بتوان اطلاعات موضعی آن را در طول زمان تحلیل کرد. با توجه به ماهیت غیرایستا بودن سیگنال‌های صوتی، تحلیل یکپارچه سیگنال ممکن است اطلاعات مهم زمانی و فرکانسی را از بین ببرد. بنابراین، پنجره‌بندی به شناسایی دقیق‌تر الگوهای زمانی و فرکانسی کمک می‌کند و امکان استخراج ویژگی‌های معنادار را فراهم می‌آورد. اندازه و نوع پنجره انتخاب‌شده در این فرآیند تأثیر مستقیمی بر کیفیت تحلیل سیگنال دارد. پنجره‌های کوتاه‌تر (معمولاً 10 تا 30 میلی‌ثانیه) برای تحلیل دقیق جزئیات زمانی مناسب هستند و برای پردازش‌هایی مانند تشخیص گفتار یا گوینده که اطلاعات زمانی بسیار مهم است، کاربرد دارند. از سوی دیگر، پنجره‌های بزرگ‌تر (30 تا 50 میلی‌ثانیه) دقت بالاتری در تحلیل فرکانسی ارائه می‌دهند و برای وظایفی مانند استخراج ویژگی‌های فرکانسی عمیق‌تر مناسب هستند. انتخاب اندازه پنجره معمولاً وابسته به کاربرد و نوع ویژگی‌هایی است که قرار است استخراج شوند. از میان انواع پنجره‌ها، پنجره همینگ (Hamming Window) و پنجره مستطیلی از پرکاربردترین گزینه‌ها هستند. پنجره همینگ با اعمال یک تابع وزن‌دهی ملایم به سیگنال، اثرات ناپیوستگی در انتهای هر بخش را کاهش می‌دهد و اطلاعات طیفی دقیق‌تری ارائه می‌دهد. این نوع پنجره در کاربردهایی که نیاز به تحلیل دقیق طیفی وجود دارد، مانند استخراج ضرایب MFCC، بسیار مؤثر است. در مقابل، پنجره مستطیلی به‌طور مستقیم سیگنال را به بخش‌های مساوی تقسیم می‌کند و اگرچه تحلیل زمانی دقیقی ارائه می‌دهد، ممکن است در تحلیل فرکانسی باعث نشت طیفی شود. یک چالش مهم در پنجره‌بندی، همپوشانی بین پنجره‌ها است. معمولاً بین 50 تا 75 درصد همپوشانی اعمال می‌شود تا اطمینان حاصل شود که اطلاعات مرزی سیگنال از دست نمی‌رود. این استراتژی به بهبود دقت تحلیل زمانی و فرکانسی کمک می‌کند و سیگنال نهایی برای تحلیل ویژگی‌هایی مانند Log Mel Spectrogram یا Spectral Centroid غنی‌تر می‌شود.

در کاربردهایی مانند تشخیص گوینده یا طبقه‌بندی جنسیت، پنجره‌بندی نه تنها امکان حفظ اطلاعات موضعی سیگنال را فراهم می‌کند، بلکه به مدل‌های یادگیری ماشین اجازه می‌دهد تا الگوهای زمانی-فرکانسی خاص مرتبط با گفتار را شناسایی کنند. این فرآیند همچنین پایه‌ای برای استخراج ویژگی‌های پیشرفته مانند Spectrograms و MFCCs محسوب می‌شود و دقت تحلیل صوتی را به‌طور چشمگیری افزایش می‌دهد. پنجره‌بندی، با مدیریت بهینه جزئیات زمانی و فرکانسی، یکی از اجزای جدایی‌ناپذیر در پیش‌پردازش صوت به شمار می‌رود.

۳_۴ اهمیت مرحله پیش‌پردازش داده‌های صوتی

اهمیت پیش‌پردازش داده‌های صوتی تا حدی است که بسیاری از متخصصان این حوزه معتقدند در صورت بی‌توجهی به آن، حتی قوی‌ترین معماری‌های شبکه عصبی نیز نمی‌توانند نتایج مطلوبی را ارائه دهند. این مرحله از چند جنبه دارای اهمیت است که در ادامه بررسی گردیده است.

۳_۴_۱ بهبود کیفیت سیگنال و حذف نویز

داده‌های صوتی اغلب در محیط‌هایی ضبط می‌شوند که دارای صداهای مزاحم مانند همهمه جمعیت، نویز تجهیزات الکترونیکی یا نویز باد هستند . این نویزها ضمن کاهش نسبت سیگنال به نویز (SNR)، جزئیات مهم صوتی را تضعیف می‌کنند و ممکن است باعث خطا در استخراج ویژگی‌های کلیدی نظیر MFCC شوند. پیش‌پردازش با استفاده از روش‌های گوناگون کاهش نویز (Noise Reduction) مانند فیلترهای باندی، موجک‌محور یا فیلترهای گابور چندمقیاسی، بخش زیادی از فرکانس‌های ناخواسته را حذف و ساختار اصلی سیگنال را حفظ می‌کند. افزون بر این، روش‌های تطبیقی مبتنی بر یادگیری عمیق قادرند ویژگی‌های نویز را به‌صورت پویا تشخیص دهند و آن را با دقت بالا حذف کنند. بدین ترتیب، پاک‌سازی سیگنال باعث می‌شود که سیستم‌های تشخیص گفتار و بازشناسی گوینده در شرایط گوناگون محیطی، از جمله محیط‌های پرنویز یا شلوغ، عملکرد با ثبات‌تری داشته باشند.

۳_۴_۲ کاهش تأثیرات دامنه و نوسان در شدت صدا

نرمال‌سازی (Normalization) فرایندی است که طی آن دامنه سیگنال به محدوده‌ای ثابت (مثلاً [1,1−][1,1-][1,1−] یا [0,1][0,1][0,1]) نگاشت می‌شود و در نتیجه از تحت‌تأثیر قرارگرفتن مدل توسط سیگنال‌های با دامنه بسیار بالا یا بسیار پایین جلوگیری می‌کند. در محیط‌های ضبط صوت واقعی، فاصله میکروفون از منبع صدا، شدت گویش فرد و دیگر عوامل محیطی می‌توانند دامنه سیگنال را به شکل قابل‌ملاحظه‌ای تغییر دهند. با اعمال نرمال‌سازی، این تفاوت‌ها همگن می‌شوند و استخراج ویژگی‌هایی نظیر MFCC و Spectral Centroid به‌صورت یکسان و بدون تحریف دامنه‌ای انجام می‌گیرد. در برخی تحقیقات از نرمال‌سازی میانگین صفر و واریانس واحد نیز استفاده می‌شود تا سیگنال در محیط آماری یکنواختی تحلیل شود.

۳_۴_۳ تمرکز بر بخش‌های معنادار سیگنال و تحلیل زمانی-فرکانسی

سیگنال‌های صوتی ماهیت غیرایستا دارند و ویژگی‌های آن‌ها در طول زمان تغییر می‌کند. بنابراین، پنجره‌بندی (Windowing) و بخش‌بندی هوشمندانه سیگنال به بازه‌های کوتاه، امکان تحلیل موضعی و شناسایی الگوهای گویش را فراهم می‌کند. انتخاب نوع پنجره (مثلاً پنجره همینگ) و میزان همپوشانی، باعث تنظیم سطح جزئیات زمانی و فرکانسی می‌شود که برای استخراج ویژگی‌هایی نظیر Log Mel Spectrogram یا MFCC ضروری است. به این ترتیب، پیش‌پردازش از طریق پنجره‌بندی و سایر تکنیک‌های مشابه، داده‌ها را برای اعمال تبدیل فوریه کوتاه‌مدت و سایر ابزارهای حوزه فرکانس آماده می‌سازد و سبب می‌شود که مدل‌های یادگیری، از محتوای زمانی-فرکانسی بهره کافی ببرند.

۳_۴_۴ استانداردسازی داده‌ها و افزایش سازگاری بین پایگاه‌های مختلف

در بسیاری از پژوهش‌ها، داده‌های صوتی از منابع و شرایط ضبط متنوعی گردآوری می‌شوند. این عدم همگونی باعث می‌شود که سیگنال‌های مشابه، توزیع آماری متفاوتی داشته باشند و مدل‌های یادگیری در مرحله آموزش دچار عدم قطعیت شوند. اعمال پیش‌پردازش‌های جامع (کاهش نویز، نرمال‌سازی، پنجره‌بندی

و غیره) به استانداردسازی داده‌ها کمک می‌کند و اثر عوامل بیرونی مانند نوع میکروفون، فاصله ضبط و شدت صدا را کاهش می‌دهد. در نتیجه، سیستم می‌تواند تمرکز خود را بر استخراج الگوهای اصلی صدا بگذارد و از اتکا به سیگنال‌های خام و نامتجانس بپرهیزد.

۳_۴_۵ ارتقای دقت و کارایی در یادگیری ماشین

از منظر طراحی سامانه‌های بلادرنگ یا محدودیت‌های محاسباتی، کاهش حجم داده پردازش‌شده و پاک‌سازی آن پیش از ورود به مدل، به‌طور محسوسی به کاهش زمان پردازش و مصرف منابع منجر می‌شود. همچنین پژوهش‌ها نشان می‌دهد اعمال پیش‌پردازش مناسب (نظیر استفاده از فیلترهای تطبیقی یا یادگیری عمیق برای حذف نویز) می‌تواند نرخ خطای شناسایی را در سیستم‌های تشخیص گفتار تا حد زیادی کاهش دهد. به‌عبارت دیگر، پیش‌پردازش نه تنها ابزاری برای بهبود سیگنال، بلکه یکی از عوامل کلیدی در افزایش دقت مدل‌های یادگیری ماشین محسوب می‌شود.

۳_۴_۶ حفاظت از ویژگی‌های ظریف صوتی

در برخی کاربردها نظیر شناسایی احساسات صوتی، تحلیل زیروبمی (Pitch) و ویژگی‌های هارمونیکی، حفظ ریزه‌کاری‌های صوتی حائز اهمیت است. روش‌های پیش‌پردازش با حذف هوشمندانه نویز و به حداقل‌رساندن تغییرات نامطلوب دامنه، اجازه می‌دهند تا سیگنال در عین تمیزشدن، الگوهای اصلی خود را از دست ندهد. این موضوع در موقعیت‌هایی که ریزترین تفاوت‌ها بین آواهای گفتار یا نوانس‌های موسیقایی نقش تعیین‌کننده دارند، بسیار ضروری است.

در مجموع، اهمیت مرحله پیش‌پردازش داده‌های صوتی به قدری است که می‌توان آن را بنیان هر نوع تحلیل و استخراج ویژگی در حوزه گفتار و موسیقی دانست. بدون طی کردن فرآیندهای اصولی در این مرحله، حتی پیشرفته‌ترین الگوریتم‌های پردازش گفتار یا طبقه‌بندی صوتی نیز با خطاهای فراوان مواجه خواهند شد. از منظر پژوهشی و صنعتی، انجام بهینه پیش‌پردازش باعث می‌شود تا سرمایه‌گذاری‌های بعدی در طراحی مدل‌های یادگیری ماشین و معماری‌های شبکه‌های عصبی به حداکثر کارایی خود دست یابند. به این ترتیب،

در تمامی گام‌های کار با سیگنال‌های صوتی، از تشخیص گفتار پیوسته گرفته تا طبقه‌بندی ژانر موسیقی و تشخیص احساس، مرحله پیش‌پردازش نقش کلیدی و غیر قابل اغماضی ایفا می‌کند و مبنای هرگونه تحلیل کارآمد و دقیق به شمار می‌رود.

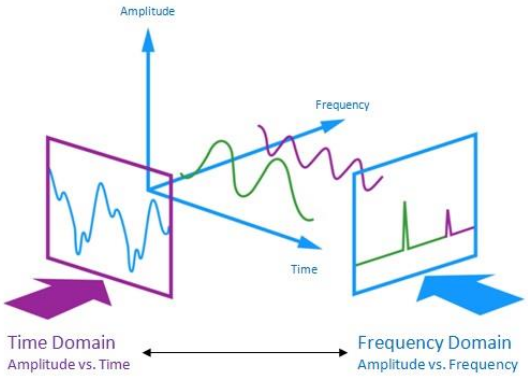
٤ تکنیک‌های استخراج ویژگی

استخراج ویژگی‌های کارآمد، یک مرحله مهم در ساخت سیستم‌های مبتنی بر یادگیری ماشین است. برای تشخیص و طبقه‌بندی صداها، باید ویژگی‌های خاصی را از صداها استخراج کنیم. دقت یک سیستم طبقه‌بندی صدا به ویژگی‌ها و روش‌های طبقه‌بندی آن بستگی دارد. نرخ نمونه‌برداری و داده‌های نمونه دو مورد اصلی هستند که یک موج صوتی از آن‌ها تشکیل شده است. با انجام چندین تبدیل بر روی این دو، می‌توان ویژگی‌های مهمی از صدا استخراج کرد. برای هر کلاس صدا، ویژگی‌هایی وجود دارد که آن را از سایر انواع صداها متمایز می‌کند. در ادامه، برخی تکنیک‌های استخراج ویژگی از فایل‌های صوتی توضیح داده شده‌اند.

۴_۱ تبدیل فوریه سریع(FFT)

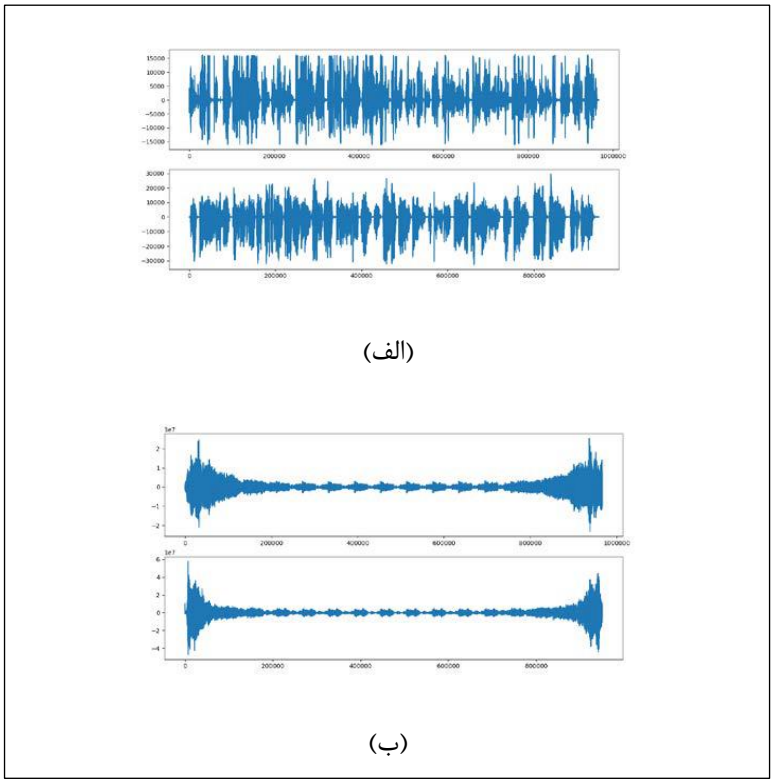
در پردازش سیگنال دیجیتال(DSP)، از تبدیل فوریه، برای تبدیل سیگنال از حوزه زمان به حوزه فرکانس استفاده می‌شود. به عبارت دیگر، با استفاده از تبدیل فوریه (FT) سیگنال را به اجزای فرکانسی‌اش تبدیل می‌کنند تا بتوان تحلیل بهتری از سیگنال داشت. این اجزا نوسانات سینوسی با فرکانس‌های مشخص هستند که هر کدام دامنه و فاز خاص خود را دارند. شکل زیر سیگنال حوزه زمان و تبدیل‌یافته آن در حوزه فرکانس را نشان می‌دهد. فرمول تبدیل فوریه گسسته نیز آورده شده است که در آن x[n] همان نمونه‌های زمانی سیگنال و N تعداد کل نمونه‌های زمانی است.

$$X[k] = \sum_{n=0}^{N-1} x[n]\, e^{-j2\pi kn/N}$$



شکل ۲: سیگنال در حوزه زمان و تبدیل یافته آن توسط تبدیل فوریه در حوزه فرکانس

در حوزه ی تشخیص و طبقه بندی صدا نیز تبدیل فوریه کار تحلیل و استخراج ویژگی‌های مهم از صدا را آسان‌تر می‌کند. در نمودارهای زیر، داده‌های صدای مرد و زن قبل و بعد از تبدیل FFT نشان داده شده‌اند. همانطور که قابل مشاهده است، سیگنال تبدیل شده بسیار قابل‌پیش‌بینی‌تر از سیگنال قبل از تبدیل است.



(الف)

(ب)

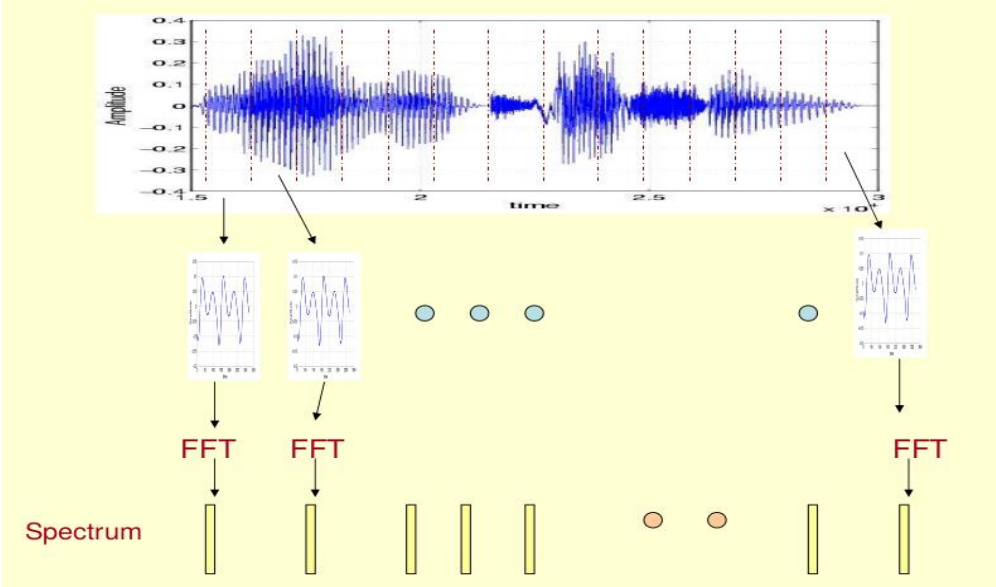
شکل ۳: الف) سیگنال صوتی مرد و زن در حوزه زمان و ب) سیگنال صوتی مرد و زن در حوزه فرکانس

معمولاً تبدیل فوریه سیگنال صوتی به‌طور مستقیم به عنوان ویژگی مورد استفاده قرار نمی‌گیرد، بلکه از آن به عنوان ورودی برای استخراج ویژگی‌های رایج در مجموعه صوتی مانند آنالیز طیفی استفاده می‌شود. برای مثال، با استفاده از تبدیل فوریه سیگنال صوتی، ویژگی‌های زیر استخراج می‌شوند.

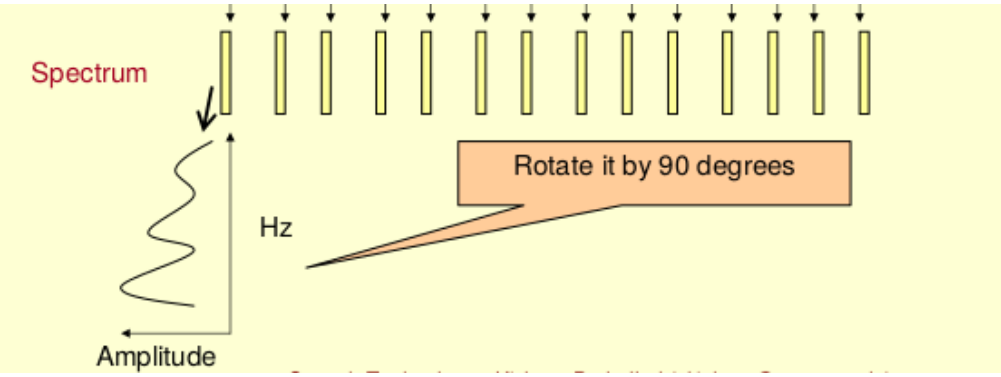
- فرکانس میانگین: فرکانس میانگین معیاری از زیر و بمی صدا است که مرکز توزیع توان در بین فرکانس‌ها را نشان می‌دهد.
- انحراف معیار: انحراف معیار مقداری است که نشان میدهد اعضای یک گروه چقدر از مقدار میانگین گروه متفاوت هستند.
- میانه: میانه مقداری است که در وسط توزیع فرکانس قرار دارد، به طوری که احتمال قرار گرفتن داده‌ها بالاتر یا پایین‌تر از آن مقدار برابر است.
- چارک سوم: چارک سوم عددی است که ۷۵٪ داده‌ها کمتر از آن عدد هستند. چارک سوم (Q75) همان میانه بخشی از داده‌ها است که بزرگتر از میانه کل داده‌ها است. این همان صدک ۷۵ است.
- چارک اول: چارک اول عددی است که ۲۵٪ داده‌ها کمتر از آن عدد هستند. این همان صدک ۲۵ است.
- دامنه بین چارکی: دامنه بین چارکی تفاوت بین چارک سوم و چارک اول است، یعنی (Q75-Q25). به نوعی پراکندگی داده‌ها را بدون نویز (داده‌های پرت) نشان می‌دهد.
- مد: مد داده‌ای است که بیشترین تکرار را در مجموعه داده‌ها دارد.

علاوه‌بر این، شکل‌های زیر نشان میدهد که چگونه با استفاده از تبدیل فوریه، می‌توان طیف‌نگار (اسپکتوگرام) مربوط به سیگنال صوتی را بدست آورد. طیف‌نگار نشان میدهد که چگونه انرژی سیگنال در فرکانس‌های مختلف توزیع شده است. برای این منظور به ازای پنجره‌های کوچکی که معمولا طول آن از نظر زمانی ۲۰ تا ۵۰ میلی‌ثانیه در نظر گرفته می‌شود، تبدیل فوریه ی سیگنال محاسبه می‌شود و با توجه به مقدار دامنه‌ی آن به ازای فرکانس‌های مختلف، یک عددی به آن نگاشته می‌شود. این عدد در واقع مقدار یک پیکسل

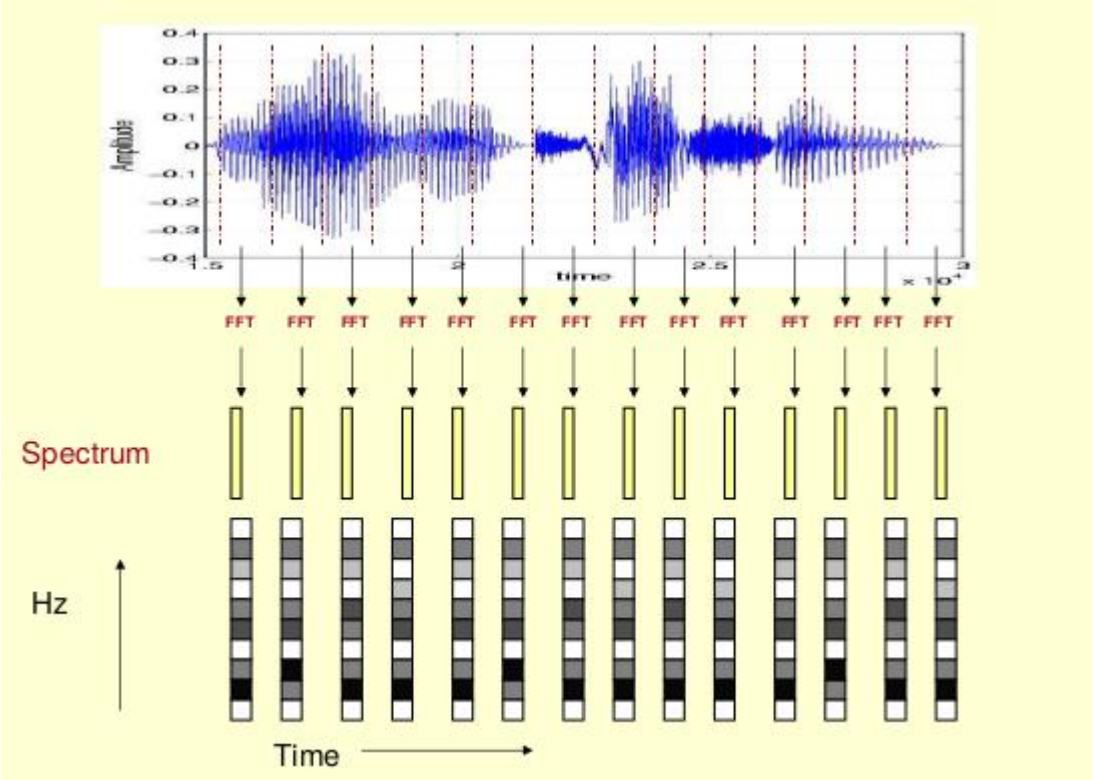
در نمایش طیف خواهد بود. محور عمودی طیف‌نگار فرکانس و محور افقی آن زمان می‌باشد که در واقع زمان مربوط به پنجره‌ای میباشد که به‌ازای آن تبدیل فوریه سیگنال در آن بازه زمانی محاسبه شده است.



شکل ۴: تقسیم سیگنال زمانی به پنجره‌های کوچکتر که با هم هم‌پوشانی دارند و محاسبه تبدیل فوریه هر یک از این پنجره‌های زمانی



شکل ۵: محور عمودی طیف‌نگار همان فرکانس است و باید مقدار عددی هر پیکسل را بر اساس اندازه دامنه در آن فرکانس برای آن طیف زمانی محاسبه کرد



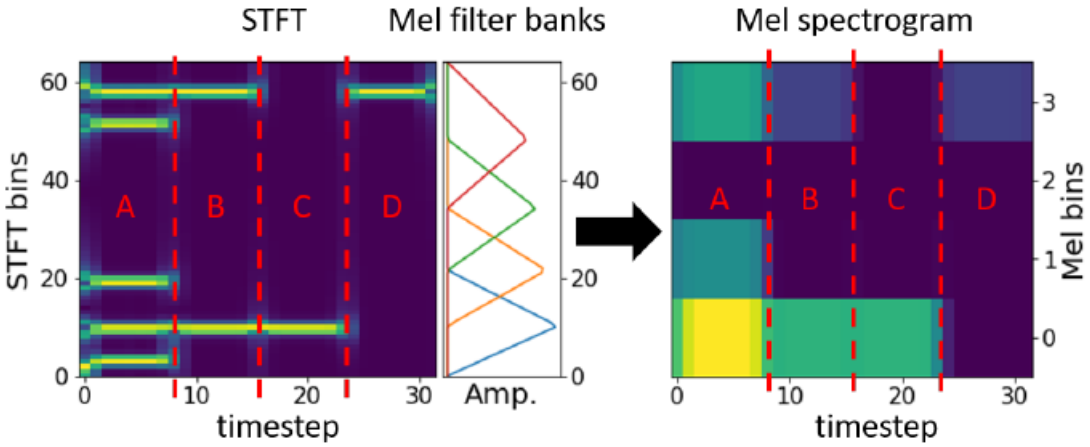
شکل ۶: طیف‌نگار مربوط به سیگنال ورودی که محور عمودی فرکانس و محور افقی زمان است

۲_۴ روش Log Mel Spectrogram

Log Mel Spectrogram یکی از ابزارهای قدرتمند در پردازش صوت است که برای بازنمایی انرژی سیگنال صوتی در دامنه فرکانس و زمان استفاده می‌شود. این تکنیک با ترکیب دو مفهوم مهم، یعنی مقیاس مل و مقیاس لگاریتمی، بازنمایی‌ای ایجاد می‌کند که با نحوه ادراک شنوایی انسان مطابقت دارد. مقیاس مل بر اساس مطالعه ادراک فرکانسی انسان طراحی شده است و هدف آن فشرده‌سازی فرکانس‌های بالا و برجسته‌سازی فرکانس‌های پایین‌تر است که برای انسان قابل تمایزتر هستند. از سوی دیگر، مقیاس لگاریتمی شدت سیگنال را به صورت غیرخطی نمایش می‌دهد، که به تقویت تغییرات کوچک در شدت صدا کمک می‌کند و موجب بهبود درک الگوهای انرژی می‌شود.

۴_۲_۱ فرآیند تولید Log Mel Spectrogram

برای تولید Log Mel Spectrogram، ابتدا سیگنال صوتی خام وارد مراحل پردازش اولیه می‌شود. این مراحل شامل نرمال‌سازی، کاهش نویز و پنجره‌بندی است. پس از آماده‌سازی سیگنال، تبدیل فوریه کوتاه‌مدت (STFT) برای انتقال سیگنال از دامنه زمانی به دامنه فرکانسی اعمال می‌شود. سپس فیلترهای مل، که مجموعه‌ای از فیلترهای باند عبور هستند، بر داده‌های فرکانسی اعمال می‌شوند. این فیلترها فرکانس‌ها را به مقیاس مل تبدیل کرده و داده‌ها را با توجه به حساسیت شنوایی انسان تنظیم می‌کنند. در نهایت، مقادیر انرژی فرکانسی محاسبه‌شده به صورت لگاریتمی مقیاس‌بندی می‌شوند، که منجر به ایجاد بازنمایی‌ای می‌شود که تغییرات شدت در بازه‌های مختلف فرکانسی را به‌طور واضح‌تری نشان می‌دهد. در شکل زیر ، نمونه‌ای از یک Mel Spectrogram نمایش داده شده است.



شکل ۷: طیفنگار مل که با ترکیب نتیجه STFT (شامل ۶۵ باند فرکانسی) با فیلتر ۴ بانکی مل به دست آمده است

۴_۲_۲ ویژگی‌های کلیدی و برتری‌ها

Log Mel Spectrogram نسبت به اسپکتروگرام معمولی برتری‌های متعددی دارد. نخستین برتری آن در هماهنگی با نحوه شنیدن انسان است؛ این بازنمایی قادر است فرکانس‌های پایین را با جزئیات بیشتری نسبت به فرکانس‌های بالا نمایش دهد، که برای کاربردهایی مانند تشخیص گفتار و طبقه‌بندی جنسیت اهمیت دارد. همچنین، مقیاس لگاریتمی در Log Mel Spectrogram باعث می‌شود اطلاعات ضعیف‌تر و

تغییرات کوچک در سیگنال، که ممکن است در اسپکتروگرام خطی نادیده گرفته شوند، برجسته شوند. این ویژگی به‌ویژه در محیط‌های نویزی بسیار کاربردی است و امکان تحلیل دقیق‌تر داده‌ها را فراهم می‌کند.

۴_۲_۳ کاربردها در سیستم‌های پردازش صوت

یکی از کاربردهای اصلی Log Mel Spectrogram در سیستم‌های تشخیص گفتار است. این بازنمایی، به دلیل توانایی آن در نمایش دقیق ویژگی‌های زمانی و فرکانسی صدا، به طور گسترده‌ای در مدل‌های یادگیری عمیق نظیر شبکه‌های عصبی کانولوشن (CNN) مورد استفاده قرار می‌گیرد. در کاربردهای تشخیص گوینده، Log Mel Spectrogram قادر است الگوهای فرکانسی و تمبروال خاص هر گوینده را بازنمایی کند، که به مدل‌های یادگیری ماشین کمک می‌کند تا با دقت بیشتری ویژگی‌های مربوط به گفتار فردی را استخراج کنند. علاوه بر این، در سیستم‌های طبقه‌بندی جنسیت، این بازنمایی توانسته است با ارائه اطلاعات دقیق درباره روشنایی صدای افراد، نتایج قابل‌اعتمادی ارائه دهد.

۴_۲_۴ تحقیقات مرتبط با Log Mel Spectrogram و تحلیل کاربردها

پژوهش‌های متعددی نشان داده‌اند که Log Mel Spectrogram به دلیل هماهنگی با نحوه ادراک انسان از صدا، عملکرد بسیار بالاتری در کاربردهای مختلف پردازش صوتی نسبت به دیگر روش‌های بازنمایی دارد. به عنوان مثال، در مطالعه‌ای که توسط Shen et al. (2024) انجام شد، استفاده از Log Mel Spectrogram در یک مدل یادگیری عمیق برای شناسایی بیماری‌های تنفسی بر اساس صداهای سرفه بررسی شد. نتایج نشان داد که این روش به دلیل توانایی آن در تفکیک دقیق انرژی فرکانس‌های پایین‌تر که معمولاً با مراحل خاصی از تنفس مرتبط هستند، عملکرد طبقه‌بندی را به‌طور قابل‌توجهی بهبود می‌بخشد.

علاوه بر این، تحقیقات Gong et al. (2022) نیز اثبات کرده‌اند که Log Mel Spectrogram در محیط‌های نویزی عملکرد بسیار بالایی دارد. در این پژوهش، نویز محیطی ابتدا با استفاده از کاهش نویز موجک حذف شد و سپس سیگنال بازنمایی‌شده با Log Mel Spectrogram به عنوان ورودی به یک مدل

طبقه‌بندی گوینده داده شد. نتایج نشان داد که این بازنمایی در تشخیص هویت افراد با نویزهای پس‌زمینه مختلف دقت بالایی ارائه می‌دهد.

۴_۲_۵ پیشرفت‌های فناورانه در استفاده از Log Mel Spectrogram

یکی از پیشرفت‌های قابل توجه در استفاده از Log Mel Spectrogram، ترکیب آن با مدل‌های یادگیری عمیق نظیر شبکه‌های عصبی کانولوشن (CNN) و شبکه‌های بازگشتی (RNN) است. این ترکیب باعث شده که مدل‌ها قادر به درک بهتر الگوهای پیچیده زمانی-فرکانسی در داده‌های صوتی شوند. به عنوان مثال، در تحقیقی که توسط Labied & Belangour (2021) انجام شد، استفاده از Log Mel Spectrogram به عنوان ورودی CNN برای تشخیص خودکار گفتار بررسی شد. این مطالعه نشان داد که مدل با دقت بالاتری نسبت به بازنمایی‌های دیگر نظیر اسپکتروگرام خطی یا ویژگی‌های زمانی-دامنه‌ای عمل می‌کند.

یکی دیگر از پیشرفت‌های جالب، استفاده از Log Mel Spectrogram به عنوان ورودی به سیستم‌های یادگیری انتقالی (Transfer Learning) است. در این سیستم‌ها، مدل‌هایی که قبلاً با داده‌های صوتی عمومی آموزش دیده‌اند، می‌توانند با تنظیمات مختصر برای وظایف خاصی نظیر طبقه‌بندی گفتار یا احساسات صوتی استفاده شوند. این رویکرد باعث کاهش زمان و منابع مورد نیاز برای آموزش مدل‌ها می‌شود.

۴_۲_۶ چالش‌ها و محدودیت‌ها

هرچند Log Mel Spectrogram یکی از پیشرفته‌ترین بازنمایی‌های صوتی است، اما با چالش‌هایی نیز مواجه است. یکی از چالش‌های اصلی، نیاز به تنظیم دقیق پارامترها نظیر اندازه پنجره، میزان همپوشانی، و تعداد فیلترهای مل است. این پارامترها مستقیماً بر کیفیت بازنمایی تأثیر می‌گذارند و باید با دقت برای هر کاربرد خاص تنظیم شوند. برای مثال، در کاربردهایی که نیاز به دقت زمانی بالا دارند، مانند تشخیص شروع و پایان کلمات، استفاده از پنجره‌های کوچک‌تر پیشنهاد می‌شود، در حالی که برای تحلیل فرکانسی بهتر، پنجره‌های بزرگ‌تر مناسب‌تر هستند.

۴_۲_۷ ارتقاء کارایی Log Mel Spectrogram در سیستم‌های یادگیری عمیق

یکی از عوامل کلیدی موفقیت Log Mel Spectrogram در پردازش صوت، ادغام آن با مدل‌های یادگیری عمیق است که توانایی استخراج الگوهای پیچیده و معنای ضمنی از داده‌های صوتی را دارند. در مطالعه‌ای که توسط Liu et al. (2024) انجام شد، Log Mel Spectrogram به عنوان ورودی به یک مدل یادگیری عمیق چندلایه با ترکیب CNN و RNN استفاده شد. این مدل توانست جزئیات زمانی-فرکانسی سیگنال را با دقت بالایی تجزیه و تحلیل کرده و برای طبقه‌بندی جنسیت و شناسایی گفتار به کار گیرد. نتایج این پژوهش نشان داد که استفاده از Log Mel Spectrogram به همراه معماری‌های پیشرفته یادگیری عمیق، دقت طبقه‌بندی را تا 95 درصد افزایش داده است.

علاوه بر این، در یک پژوهش دیگر توسط Shintri & Bhatia (2015)، Log Mel Spectrogram به عنوان ورودی به یک شبکه کانولوشنی چندسطحی (Deep CNN) مورد بررسی قرار گرفت. این پژوهش تمرکز بر تشخیص هیجانات صوتی داشت و نشان داد که با استفاده از Log Mel Spectrogram، دقت در تشخیص هیجانات حتی در محیط‌های نویزی به طور چشمگیری افزایش یافت. این عملکرد بالا ناشی از توانایی Log Mel Spectrogram در حفظ ویژگی‌های مهم صوتی و کاهش تأثیر نویزهای پس‌زمینه است.

۴_۲_۸ تطبیق Log Mel Spectrogram برای کاربردهای چندگانه

یکی از ویژگی‌های مهم Log Mel Spectrogram، تطبیق‌پذیری آن برای کاربردهای مختلف است. این بازنمایی می‌تواند برای وظایفی چون شناسایی موسیقی، تشخیص احساسات، و حتی تحلیل صداهای محیطی استفاده شود. برای مثال، در پژوهشی که توسط Labied & Belangour (2021) انجام شد، از Log Mel Spectrogram برای طبقه‌بندی صداهای محیطی استفاده شد و دقت 92 درصدی در تشخیص صداهایی مانند باران، صدای ماشین، و گفتگو به دست آمد. این نتایج نشان‌دهنده توانایی این بازنمایی در تفکیک فرکانس‌های مختلف و تحلیل جزئیات زمانی سیگنال است.

علاوه بر این، در مطالعه‌ای که توسط Gong et al. (2022) انجام شد، از Log Mel Spectrogram برای تشخیص بیماری‌های تنفسی بر اساس صداهای سرفه استفاده شد. این پژوهش نشان داد که بازنمایی مل به دلیل توانایی آن در تفکیک انرژی فرکانس‌های مرتبط با مراحل مختلف سرفه (مانند شروع، اوج و پایان)، اطلاعات دقیقی برای تحلیل و طبقه‌بندی فراهم می‌کند. این ویژگی به‌ویژه در کاربردهایی که سیگنال‌های پیچیده و چندمرحله‌ای دارند، بسیار حیاتی است.

۴_۲_۹ نوآوری‌ها در طراحی و تنظیم پارامترهای Log Mel Spectrogram

پیشرفت‌های اخیر در تنظیم پارامترهای Log Mel Spectrogram باعث بهبود چشمگیر کارایی آن شده است. انتخاب تعداد مناسب فیلترهای مل، تنظیم طول پنجره و میزان همپوشانی از جمله پارامترهایی هستند که به‌طور مستقیم بر کیفیت بازنمایی تأثیر می‌گذارند. برای مثال، در پژوهش Makarem (2023)، اثرات تغییر تعداد فیلترهای مل بر دقت تشخیص مورد بررسی قرار گرفت. نتایج نشان داد که با افزایش تعداد فیلترها، دقت بازنمایی فرکانسی افزایش می‌یابد، اما این افزایش ممکن است به هزینه زمان پردازش بیشتر منجر شود. این یافته‌ها به اهمیت تنظیم دقیق پارامترها متناسب با نیازهای هر کاربرد اشاره دارند.

۴_۲_۱۰ استفاده از Log Mel Spectrogram در کاربردهای بلادرنگ

یکی از زمینه‌های جذاب استفاده از Log Mel Spectrogram، کاربردهای بلادرنگ (Real-Time Applications) است. برای مثال، در سیستم‌های تشخیص گفتار و شناسایی گوینده که نیاز به پردازش سریع دارند، Log Mel Spectrogram به دلیل بازنمایی فشرده و دقیق خود بسیار موثر عمل می‌کند. در پژوهشی توسط Hashim & Karam (2021)، از Log Mel Spectrogram برای تشخیص گوینده در یک سیستم بلادرنگ استفاده شد. نتایج این پژوهش نشان داد که این بازنمایی به دلیل زمان پردازش کوتاه‌تر و دقت بالا، برای سیستم‌های بلادرنگ بسیار مناسب است.

۴_۲_۱۱ چالش‌های استفاده از Log Mel Spectrogram و راهکارهای پیشنهادی

در حالی که Log Mel Spectrogram به‌عنوان یک ابزار قدرتمند در پردازش صوت شناخته می‌شود، چالش‌هایی نیز در استفاده از آن وجود دارد. یکی از این چالش‌ها، حساسیت به تغییرات محیطی و شرایط ضبط صوت است. برای مثال، کیفیت میکروفون، فاصله گوینده از دستگاه ضبط، و نویز پس‌زمینه می‌تواند بر دقت ویژگی‌های استخراج‌شده از Log Mel Spectrogram تأثیر بگذارد. این مشکل در سیستم‌هایی که در محیط‌های پرنویز کار می‌کنند، مانند سیستم‌های شناسایی گوینده در مراکز تماس، به طور خاص برجسته است. برای مقابله با این چالش، استفاده از تکنیک‌هایی نظیر کاهش نویز پیشرفته، مانند فیلترهای انطباقی و تکنیک‌های مبتنی بر یادگیری عمیق، پیشنهاد شده است. این روش‌ها به پاکسازی سیگنال ورودی و حذف نویزهای مزاحم کمک می‌کنند.

چالش دیگر، نیاز به محاسبات سنگین در کاربردهای بلادرنگ است. محاسبه Log Mel Spectrogram، به‌ویژه برای سیگنال‌های صوتی طولانی یا جریان‌های صوتی زنده، می‌تواند زمان‌بر باشد و منابع محاسباتی زیادی را مصرف کند. برای حل این مشکل، روش‌هایی مانند استفاده از فیلترهای مل از پیش‌محاسبه‌شده یا بهینه‌سازی الگوریتم تبدیل فوریه کوتاه‌مدت (STFT) پیشنهاد شده است. این تکنیک‌ها می‌توانند سرعت پردازش را افزایش دهند بدون آنکه دقت بازنمایی کاهش یابد.

یکی دیگر از چالش‌ها، تنظیم دقیق پارامترهای Log Mel Spectrogram است. انتخاب تعداد فیلترهای مل، اندازه پنجره و میزان همپوشانی می‌تواند تأثیرات متناقضی بر دقت و زمان پردازش داشته باشد. برای مثال، در مطالعه‌ای که توسط Liu et al. (2024) انجام شد، مشخص شد که تعداد زیاد فیلترهای مل می‌تواند جزئیات بیشتری را به‌دست آورد، اما ممکن است باعث افزایش پیچیدگی محاسبات شود. استفاده از الگوریتم‌های بهینه‌سازی، مانند روش‌های مبتنی بر جستجوی شبکه‌ای (Grid Search) یا بهینه‌سازی بیزی (Bayesian Optimization)، می‌تواند به یافتن تنظیمات بهینه کمک کند.

۴_۲_۱۲ تطبیق Log Mel Spectrogram با کاربردهای خاص

یکی از مزیت‌های قابل توجه Log Mel Spectrogram، انعطاف‌پذیری آن برای تنظیم متناسب با کاربردهای خاص است. برای مثال، در کاربردهای تشخیص ژانر موسیقی، تمرکز بر فرکانس‌های پایین‌تر ممکن است به تفکیک بهتر ژانرها کمک کند، زیرا بسیاری از ژانرهای موسیقی دارای الگوهای فرکانسی خاص در محدوده فرکانس‌های پایین هستند. در مقابل، در کاربردهای تشخیص گوینده، اطلاعات موجود در فرکانس‌های میانی و بالاتر می‌تواند اطلاعات مهم‌تری درباره تمایز گویندگان ارائه دهد.

برای بهبود تطبیق با کاربردهای خاص، ترکیب Log Mel Spectrogram با ویژگی‌های تکمیلی نظیر Spectral Centroid و MFCC می‌تواند نتایج بهتری ارائه دهد. برای مثال، در مطالعه‌ای که توسط Labied & Belangour (2021) انجام شد، ترکیب این ویژگی‌ها برای تشخیص گفتار استفاده شد و دقت سیستم را در مقایسه با استفاده از Log Mel Spectrogram به‌تنهایی بهبود بخشید. این رویکرد نشان‌دهنده قدرت ترکیب بازنمایی‌ها برای تقویت توانایی مدل‌های یادگیری ماشین است.

تحقیقات اخیر نشان داده‌اند که Log Mel Spectrogram، در کنار سایر بازنمایی‌های صوتی، به یکی از اصلی‌ترین ابزارهای پردازش صوت در سیستم‌های مدرن تبدیل شده است. با پیشرفت در سخت‌افزارهای محاسباتی و الگوریتم‌های بهینه‌سازی، انتظار می‌رود که این بازنمایی در کاربردهای بلادرنگ نیز محبوب‌تر شود. همچنین، ترکیب Log Mel Spectrogram با مدل‌های پیشرفته یادگیری عمیق، مانند مدل‌های Transformer، می‌تواند قابلیت‌های جدیدی برای تحلیل و طبقه‌بندی صوت فراهم کند. در نهایت، پژوهش‌های آینده باید به بهبود تحمل‌پذیری Log Mel Spectrogram نسبت به نویز و تغییرات محیطی، کاهش زمان پردازش، و توسعه روش‌های نوین برای ترکیب آن با ویژگی‌های دیگر متمرکز شوند تا بتوانند نتایج دقیق‌تر و کارآمدتری ارائه دهند.

۴_۳ روش MFCC

Mel Frequency Cepstral Coefficients (MFCC) یکی از بنیادی‌ترین تکنیک‌ها در پردازش سیگنال صوتی است که در بسیاری از کاربردهای کلیدی همچون تشخیص گفتار، شناسایی گوینده، تحلیل موسیقی و حتی تشخیص احساسات به کار می‌رود. MFCC نه‌تنها به دلیل سادگی و سرعت، بلکه به خاطر تطبیق با نحوه شنیداری انسان، به یک استاندارد در تحلیل صوت تبدیل شده است. این تکنیک با استفاده از مفاهیمی از شنوایی انسانی، سیگنال صوتی را به ویژگی‌های فشرده و قابل استفاده برای مدل‌های یادگیری ماشین تبدیل می‌کند. به طور خاص، MFCC با تمرکز بر روی فرکانس‌های پایین، که در تحلیل گفتار و موسیقی بیشتر اهمیت دارند، مزایای بزرگی نسبت به تکنیک‌های سنتی دارد.

MFCC همچنین به دلیل توانایی خود در کاهش حساسیت به تغییرات دامنه و نویز، برای کاربردهایی در محیط‌های چالش‌برانگیز مانند محیط‌های نویزی، تماس‌های تلفنی یا ضبط‌های با کیفیت پایین، مورد توجه ویژه قرار گرفته است. یکی از دلایل موفقیت این روش، توانایی آن در تبدیل اطلاعات زمانی-فرکانسی به یک نمایش فشرده است که امکان تحلیل سریع و دقیق را فراهم می‌آورد. علاوه بر این، MFCC به دلیل قابلیت کاهش بعد، امکان پردازش حجم بالای داده‌های صوتی را با کاهش نیاز به منابع محاسباتی میسر می‌سازد.

۴_۳_۱ فرآیند تولید ضرایب MFCC

تولید MFCC شامل چندین مرحله کلیدی است که هر کدام نقش مهمی در بازنمایی سیگنال صوتی ایفا می‌کنند. این مراحل عبارتند از:

۱) پنجره‌بندی سیگنال صوتی: سیگنال صوتی خام به پنجره‌های زمانی کوچک تقسیم می‌شود، زیرا سیگنال‌های صوتی معمولاً غیرایستا هستند و نیاز به تجزیه به بخش‌های کوچک‌تر دارند. این پنجره‌ها معمولاً 20 تا 40 میلی‌ثانیه طول دارند و همپوشانی حدود 50 درصدی بین پنجره‌ها اعمال می‌شود. این گام به تحلیل ویژگی‌های موضعی سیگنال در طول زمان کمک می‌کند.

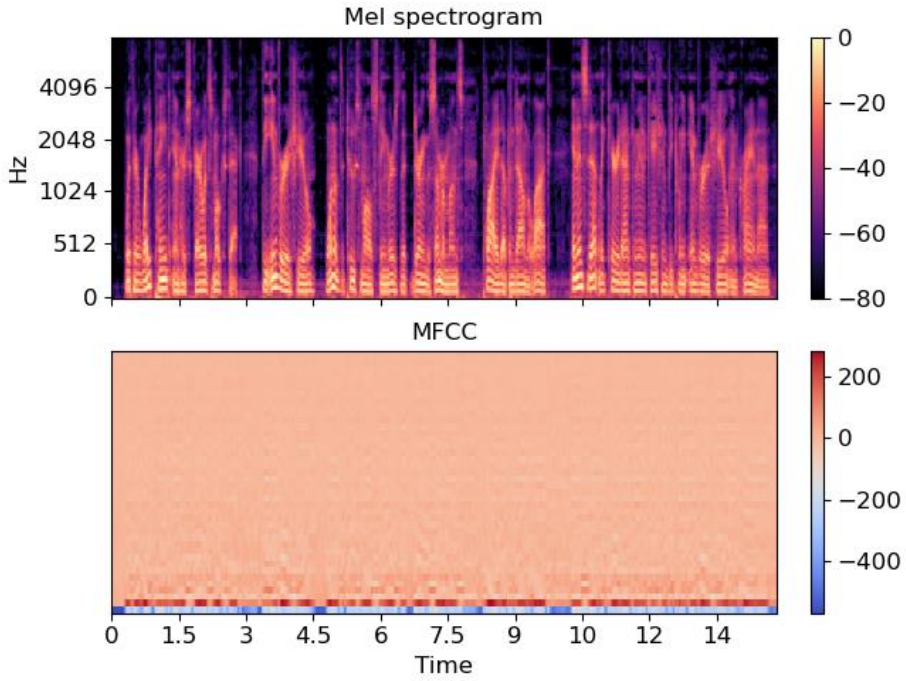
۲) اعمال تبدیل فوریه کوتاه‌مدت (STFT): در این مرحله، هر پنجره زمانی با استفاده از تبدیل فوریه کوتاه‌مدت از حوزه زمان به حوزه فرکانس تبدیل می‌شود. این تبدیل، توزیع انرژی فرکانسی سیگنال را در طول زمان نشان می‌دهد و پایه‌ای برای محاسبات بعدی است. این مرحله به شناسایی فرکانس‌های غالب سیگنال کمک می‌کند و نمایش دقیقی از الگوهای فرکانسی ارائه می‌دهد.

۳) فیلترگذاری به کمک مقیاس مل: در این مرحله، مقیاس مل برای تطبیق بهتر با نحوه ادراک انسان از فرکانس اعمال می‌شود. این مقیاس فرکانس‌های پایین را با دقت بیشتری نسبت به فرکانس‌های بالا تحلیل می‌کند. فیلترهای مل، معمولاً به صورت یک بانک فیلتر مثلثی تعریف می‌شوند، که هر کدام بخشی از طیف فرکانسی را پوشش می‌دهند. این مرحله تضمین می‌کند که اطلاعات مرتبط با درک صوتی حفظ شود.

۴) محاسبه انرژی لگاریتمی: انرژی خروجی از هر فیلتر مثلثی، به صورت لگاریتمی محاسبه می‌شود. این گام به تقویت سیگنال‌های ضعیف‌تر کمک می‌کند و در عین حال، از تاثیر غیرخطی فرکانس‌های قوی جلوگیری می‌کند. این ویژگی باعث می‌شود MFCC به تغییرات شدت سیگنال مقاوم باشد.

۵) اعمال تبدیل کسینوسی گسسته (DCT): در نهایت، برای کاهش ابعاد و حذف همبستگی بین خروجی‌های فیلتر مل، تبدیل کسینوسی گسسته اعمال می‌شود. این تبدیل داده‌ها را به ضرایب MFCC تبدیل می‌کند که معمولاً بین 12 تا 20 ضریب اولیه برای هر پنجره زمانی انتخاب می‌شوند. این ضرایب نمایانگر ویژگی‌های کلیدی سیگنال صوتی هستند که می‌توانند برای تحلیل‌های بعدی استفاده شوند.

در شکل زیر MFCC و Spectrogram برای یک سیگنال صوتی نمایش داده شده‌اند.



شکل ۸: MFCC و Spectrogram برای یک سیگنال صوتی

یکی از بزرگ‌ترین مزایای MFCC، توانایی آن در ارائه بازنمایی فشرده از سیگنال صوتی است که شامل ویژگی‌های شنیداری معنادار می‌باشد. این بازنمایی به ویژه در مدل‌های یادگیری ماشین، جایی که حجم داده‌ها و منابع محاسباتی محدود هستند، بسیار کاربردی است. به علاوه، MFCC با حذف اطلاعات غیرضروری و تمرکز بر فرکانس‌های کلیدی، دقت تشخیص را به‌طور چشمگیری افزایش می‌دهد MFCC.همچنین به دلیل انعطاف‌پذیری خود، در بسیاری از زمینه‌ها از جمله تشخیص گفتار، شناسایی گوینده، طبقه‌بندی جنسیت و تحلیل موسیقی استفاده می‌شود. برای مثال، در سیستم‌های تشخیص گفتار، MFCC اطلاعات مهم مربوط به الگوهای فرکانسی خاص گفتار را استخراج می‌کند که برای تفکیک کلمات و جملات ضروری است. در حوزه شناسایی گوینده، MFCC به دلیل توانایی در استخراج الگوهای فرکانسی خاص هر فرد، به یکی از ابزارهای اصلی تبدیل شده است. ویژگی‌هایی نظیر فرکانس‌های هارمونیک و الگوهای فرکانسی بیان، که در ضرایب MFCC برجسته می‌شوند، امکان تفکیک گویندگان مختلف را فراهم می‌کنند. این ویژگی‌ها در ترکیب با تکنیک‌های یادگیری عمیق مانند شبکه‌های عصبی بازگشتی (RNN) و ماشین‌های بردار پشتیبانی (SVM) می‌توانند دقت تشخیص گوینده را به میزان قابل‌توجهی افزایش دهند. برای مثال، در محیط‌های نویزی، ترکیب MFCC با روش‌های پیش‌پردازشی مانند کاهش نویز موجک و نرمال‌سازی، عملکرد سیستم را بهبود می‌بخشد.

در کاربردهای موسیقی، MFCC به عنوان ابزاری برای تحلیل تمبروال و طبقه‌بندی ژانر موسیقی به کار می‌رود. ضرایب MFCC اطلاعاتی را ارائه می‌دهند که ویژگی‌های صوتی مانند روشنایی یا تیرگی صدا را مشخص می‌کنند. این اطلاعات در ترکیب با مدل‌های یادگیری ماشین، امکان شناسایی دقیق ژانرهای موسیقی، آهنگ‌سازها و حتی احساسات موجود در قطعات موسیقی را فراهم می‌کند.

پیشرفت‌های اخیر در تکنیک‌های MFCC، بهبودهایی را در دقت و انعطاف‌پذیری این روش ایجاد کرده‌اند. یکی از این پیشرفت‌ها، ترکیب MFCC با تکنیک‌های چندتابعی (Multitaper) است. این روش از چندین پنجره برای تحلیل سیگنال استفاده می‌کند و با کاهش واریانس تخمین طیف فرکانسی، دقت بیشتری در استخراج ویژگی‌ها ارائه می‌دهد. این تکنیک به‌ویژه در محیط‌های نویزی یا در تحلیل داده‌های صوتی با تنوع بالا بسیار موثر است.

تکنیک دیگری که به توسعه MFCC کمک کرده است، استفاده از ضرایب گاماتون (Gammatone Filters) است. این فیلترها که بر اساس مدل‌های شنوایی انسان طراحی شده‌اند، با شبیه‌سازی پاسخ غشای باسیلار، اطلاعات دقیق‌تری درباره فرکانس‌های کلیدی سیگنال صوتی ارائه می‌دهند. ضرایب گاماتون به‌طور خاص در کاربردهایی نظیر تشخیص گفتار و گوینده که به دقت بالاتری نیاز دارند، مفید هستند.

۴_۳_۲ ادغام MFCC با مدل‌های یادگیری عمیق

ترکیب MFCC با مدل‌های یادگیری عمیق مانند شبکه‌های عصبی پیچشی (CNN) و شبکه‌های بازگشتی (RNN)، به‌طور قابل‌توجهی توانایی این تکنیک در تحلیل داده‌های صوتی را افزایش داده است. این مدل‌ها با توانایی در یادگیری الگوهای پیچیده، می‌توانند از ضرایب MFCC برای شناسایی ویژگی‌های غیرخطی و پنهان سیگنال استفاده کنند. به عنوان مثال، در تشخیص احساسات صوتی، ترکیب MFCC با مدل‌های CNN توانسته است الگوهای عاطفی موجود در صدای انسان را با دقت بیشتری استخراج کند.

۴_۳_۳ چالش‌ها و راه‌حل‌ها در استفاده از MFCC

اگرچه MFCC یکی از قدرتمندترین ابزارها در پردازش صوت است، همچنان با چالش‌هایی مواجه است. یکی از چالش‌ها، حساسیت MFCC به نویزهای محیطی است. سیگنال‌های نویزی می‌توانند اطلاعات فرکانسی مهم را تحت‌الشعاع قرار داده و دقت ضرایب MFCC را کاهش دهند. برای رفع این چالش، روش‌هایی مانند کاهش نویز تطبیقی، استفاده از فیلترهای پیشرفته مانند Wiener Filters و ترکیب MFCC با تکنیک‌های حذف نویز، اثربخشی بیشتری ارائه می‌دهند.

چالش دیگر، انتخاب تعداد مناسب ضرایب MFCC برای هر کاربرد است. اگر تعداد ضرایب کم باشد، ممکن است اطلاعات کلیدی از دست برود و اگر تعداد زیاد باشد، مدل به پیچیدگی محاسباتی بیشتری نیاز دارد. استفاده از روش‌های بهینه‌سازی مانند Cross-Validation یا الگوریتم‌های انتخاب ویژگی می‌تواند به تنظیم تعداد ضرایب MFCC کمک کند.

۴_۳_۴ موارد استفاده در حوزه‌های متنوع

MFCC به دلیل انعطاف‌پذیری و تطبیق با طیف گسترده‌ای از کاربردها در تحلیل صوت، یکی از محبوب‌ترین تکنیک‌های بازنمایی صوت محسوب می‌شود. در سیستم‌های تشخیص گفتار، MFCC با استخراج الگوهای فرکانسی مرتبط با صداهای خاص گفتاری، به بازشناسی دقیق کلمات و جملات کمک می‌کند. این تکنیک به‌ویژه در سیستم‌های چندزبانه کاربرد دارد، زیرا ضرایب MFCC می‌توانند ویژگی‌های خاص زبان‌ها یا لهجه‌های مختلف را با دقت بالا بازنمایی کنند. برای مثال، در تشخیص کلمات کلیدی در صداهای نویزی، ترکیب MFCC با مدل‌های عصبی بازگشتی (RNN) یا شبکه‌های عصبی پیچشی (CNN)، عملکرد بسیار موفقی از خود نشان داده است.

در حوزه‌های پزشکی، MFCC برای تحلیل صداهای حیاتی مانند صدای تنفس، سرفه و صدای قلب استفاده می‌شود. این ویژگی‌ها به پزشکان در تشخیص بیماری‌هایی مانند آسم، بیماری‌های ریوی و حتی

مشکلات قلبی کمک می‌کنند. ضرایب MFCC با نمایش دقیق الگوهای فرکانسی خاص این صداها، امکان شناسایی تغییرات کوچک و نشانه‌های اولیه بیماری را فراهم می‌آورند.

در موسیقی و تحلیل صوتی، MFCC برای استخراج ویژگی‌هایی مانند تمبروال صدا، طبقه‌بندی ژانر موسیقی و شناسایی سازها به کار می‌رود. این ضرایب، اطلاعاتی درباره روشنایی یا تیرگی صدای موسیقی ارائه می‌دهند که برای تحلیل دقیق‌تر آهنگ‌ها و قطعات موسیقی مفید است. علاوه بر این، در سیستم‌های توصیه موسیقی، ضرایب MFCC می‌توانند به شناسایی ترجیحات شنیداری کاربران کمک کرده و پیشنهادات مرتبط‌تری ارائه دهند.

۴_۳_۵ پیشرفت‌های تکنولوژیک در بهبود MFCC

پیشرفت‌های اخیر در تکنولوژی، قابلیت‌های MFCC را به سطح بالاتری ارتقا داده است. یکی از این پیشرفت‌ها، استفاده از تکنیک‌های پردازش سیگنال پیشرفته مانند تبدیل موجک (Wavelet Transform) در کنار ضرایب MFCC است. این ترکیب به شناسایی ویژگی‌های زمانی-فرکانسی دقیق‌تر کمک کرده و عملکرد سیستم‌های تشخیص گفتار و موسیقی را بهبود بخشیده است.

علاوه بر این، ترکیب MFCC با تکنیک‌های مدل‌سازی شنوایی، مانند استفاده از فیلترهای گاماتون، امکان استخراج ویژگی‌های شبیه‌سازی‌شده از سیستم شنوایی انسان را فراهم کرده است. این ویژگی‌ها به‌ویژه در سیستم‌های شناسایی گوینده و تشخیص احساسات، مزیت‌های چشمگیری به همراه داشته‌اند.

در کل ،MFCC یکی از ابزارهای بنیادی و قدرتمند در تحلیل صوت است که به دلیل ویژگی‌های منحصربه‌فردش، در بسیاری از حوزه‌های علمی و صنعتی مورد استفاده قرار می‌گیرد. توانایی این تکنیک در استخراج ویژگی‌های کلیدی، کاهش بعد و مقاومت در برابر نویز، آن را به انتخابی ایده‌آل برای سیستم‌های پردازش صوتی تبدیل کرده است. پیشرفت‌های اخیر در ترکیب MFCC با تکنیک‌های یادگیری ماشین و پردازش سیگنال، پتانسیل این ابزار را برای کاربردهای آینده تقویت کرده و راه را برای بهبود بیشتر سیستم‌های پردازش صوت هموار ساخته است.

این انعطاف‌پذیری و قابلیت تطبیق MFCC با شرایط متنوع، از کاربردهای ساده تشخیص گفتار تا تحلیل‌های پیچیده پزشکی و موسیقی، نشان‌دهنده اهمیت بی‌چون‌وچرای این تکنیک در حوزه پردازش صوت است.

۴_۳_۶ نقش کلیدی در سیستم‌های صوتی پیشرفته

Mel Frequency Cepstral Coefficients (MFCC) در دهه‌های اخیر به یکی از ابزارهای اصلی پردازش صوتی تبدیل شده است، به دلیل توانایی در استخراج ویژگی‌های مهم شنیداری از سیگنال‌های صوتی. این تکنیک با استفاده از مفاهیم شنوایی انسان، الگوهای فرکانسی مهم را برجسته می‌کند و در عین حال، اطلاعات غیرضروری یا نویز را کاهش می‌دهد. MFCC به‌ویژه در محیط‌های نویزی یا کاربردهایی که داده‌های صوتی خام پیچیده و متنوع هستند، عملکرد بسیار موثری دارد.

فرآیند استخراج ضرایب MFCC شامل چند مرحله کلیدی است. ابتدا، سیگنال صوتی خام با استفاده از پنجره‌بندی به بخش‌های کوچکتر تقسیم می‌شود. این پنجره‌ها معمولاً بین 20 تا 40 میلی‌ثانیه طول دارند و همپوشانی 50 درصدی برای اطمینان از حفظ اطلاعات مرزی اعمال می‌شود. سپس، هر پنجره به کمک تبدیل فوریه کوتاه‌مدت (STFT) به حوزه فرکانس منتقل می‌شود. این تبدیل امکان شناسایی الگوهای فرکانسی سیگنال را فراهم می‌کند و پایه‌ای برای اعمال فیلتر مل می‌سازد. فیلتر مل، که مقیاسی بر اساس نحوه ادراک انسان از فرکانس‌ها است، به شناسایی بهتر فرکانس‌های پایین که برای تحلیل گفتار مهم‌تر هستند، کمک می‌کند. در نهایت، تبدیل کسینوسی گسسته (DCT) برای کاهش بعد و حذف همبستگی بین ضرایب اعمال می‌شود.

MFCC به‌ویژه در سیستم‌های تشخیص گفتار به‌عنوان استاندارد پذیرفته شده است. این تکنیک ویژگی‌های صوتی مرتبط با حروف، کلمات و جملات را استخراج می‌کند که برای بازشناسی دقیق گفتار ضروری است. علاوه بر این، MFCC در سیستم‌های شناسایی گوینده به دلیل توانایی در استخراج الگوهای خاص فرکانسی که به هر گوینده منحصربه‌فرد است، کاربرد گسترده‌ای دارد. در این زمینه، ترکیب MFCC با

مدل‌های یادگیری ماشین، مانند ماشین‌های بردار پشتیبانی (SVM) و شبکه‌های عصبی بازگشتی (RNN)، عملکرد سیستم‌ها را به میزان قابل‌توجهی بهبود بخشیده است.

در موسیقی، MFCC برای تحلیل تمبروال، طبقه‌بندی ژانرها، و شناسایی سازها استفاده می‌شود. ضرایب MFCC اطلاعات مربوط به روشنایی یا تیرگی صدا را ارائه می‌دهند که برای تحلیل موسیقی و ایجاد سیستم‌های توصیه موسیقی حیاتی است. به‌عنوان مثال، در طبقه‌بندی ژانر موسیقی، MFCC می‌تواند ویژگی‌های خاص ژانرهای مختلف را بازنمایی کند، که به تشخیص دقیق‌تر کمک می‌کند.

چالش‌های استفاده از MFCC شامل حساسیت آن به نویز و تغییرات شرایط ضبط است. برای مقابله با این چالش‌ها، روش‌های پیشرفته مانند ترکیب MFCC با کاهش نویز موجک یا استفاده از ضرایب گاماتون توسعه یافته‌اند. این روش‌ها به بهبود عملکرد MFCC در محیط‌های نویزی و شرایط پیچیده کمک کرده‌اند. علاوه بر این، تکنیک‌های مدرن مانند Multitaper MFCC، که از چندین پنجره برای تحلیل دقیق‌تر استفاده می‌کند، دقت و انعطاف‌پذیری این تکنیک را افزایش داده‌اند.

به طور خلاصه، MFCC به دلیل قابلیت‌های منحصر به فرد خود در استخراج و فشرده‌سازی ویژگی‌های شنیداری، یکی از ابزارهای غیرقابل جایگزین در پردازش صوت است. پیشرفت‌های اخیر در تکنیک‌های پردازش سیگنال و ادغام MFCC با مدل‌های یادگیری عمیق، آینده‌ای روشن برای این ابزار در کاربردهای متنوع پردازش صوتی به تصویر می‌کشد.

۴_۴ روش Spectral Centroid

Spectral Centroid به عنوان یکی از ویژگی‌های اساسی طیفی در پردازش صوت، اطلاعاتی حیاتی در مورد توزیع انرژی فرکانسی یک سیگنال ارائه می‌دهد. این ویژگی، به‌طور ساده، به معنای تعیین مرکز ثقل انرژی در طیف فرکانسی است. Spectral Centroid نقش کلیدی در بازنمایی تمبروال سیگنال‌های صوتی ایفا می‌کند و به دلیل قابلیت‌های منحصربه‌فرد در تحلیل توزیع انرژی صوتی، به یکی از ابزارهای پرکاربرد در تحلیل و پردازش سیگنال‌های صوتی تبدیل شده است. این ویژگی در طیف وسیعی از کاربردها، از موسیقی و

گفتار گرفته تا تحلیل احساسات صوتی، اهمیت دارد و به دلیل سادگی محاسبات و ارتباط قوی با ادراک شنوایی انسان، بسیار مورد توجه قرار گرفته است.

در زمینه موسیقی، Spectral Centroid به‌عنوان شاخصی برای تحلیل روشنایی یا تیرگی صدا مورد استفاده قرار می‌گیرد. در این کاربرد، روشنایی صدا اغلب با فرکانس‌های بالا و مقادیر بالاتر Spectral Centroid مرتبط است. به‌عنوان مثال، سازهایی مانند ویولن یا پیانو، که صدای شفاف و درخشانی دارند، مقدار بالایی از Spectral Centroid تولید می‌کنند. از سوی دیگر، سازهایی مانند کنترباس یا سازهای کوبه‌ای سنگین، به دلیل غالب بودن فرکانس‌های پایین، مقدار کمتری از این ویژگی را نشان می‌دهند. این ویژگی به طور خاص در طبقه‌بندی ژانرهای موسیقی نیز مورد استفاده قرار می‌گیرد، چرا که ژانرهایی مانند موسیقی الکترونیک یا پاپ تمایل به مقادیر بالاتر Spectral Centroid دارند، در حالی که ژانرهایی مانند جاز یا کلاسیک معمولاً مقادیر پایین‌تری دارند.

در حوزه گفتار و پردازش صوت انسانی، Spectral Centroid به‌عنوان معیاری برای تحلیل ویژگی‌های جنسیتی و همچنین تفکیک گوینده استفاده می‌شود. صداهای زنان به دلیل فرکانس‌های بالاتر به‌طور معمول مقدار بالاتری از Spectral Centroid نسبت به صداهای مردان دارند. این تمایز فرکانسی به الگوریتم‌های پردازش گفتار امکان می‌دهد که با استفاده از Spectral Centroid به عنوان یک ویژگی کلیدی، جنسیت گوینده را با دقت بالایی تشخیص دهند. علاوه بر این، این ویژگی در تحلیل احساسات صوتی نیز نقش مهمی دارد. احساسات شادی و هیجان معمولاً با صداهای روشن‌تر و مقادیر بالاتر Spectral Centroid همراه هستند، در حالی که احساسات غم و خشم به فرکانس‌های پایین‌تر و مقادیر کمتر این ویژگی نسبت داده می‌شوند.

چالش اصلی در استفاده از Spectral Centroid، حساسیت آن به نویز است. نویزهای محیطی، به‌ویژه نویزهای دارای فرکانس بالا، می‌توانند مقادیر Spectral Centroid را به صورت نادرست افزایش داده و دقت تحلیل را کاهش دهند. این مسئله در محیط‌های شلوغ یا ضبط‌های با کیفیت پایین به وضوح مشهود است. برای مقابله با این چالش، تکنیک‌های پیش‌پردازش نظیر کاهش نویز با استفاده از فیلترهای تطبیقی، کاهش نویز موجک، و فیلترهای باندی مورد استفاده قرار می‌گیرند. این روش‌ها نه تنها به بهبود کیفیت سیگنال خام

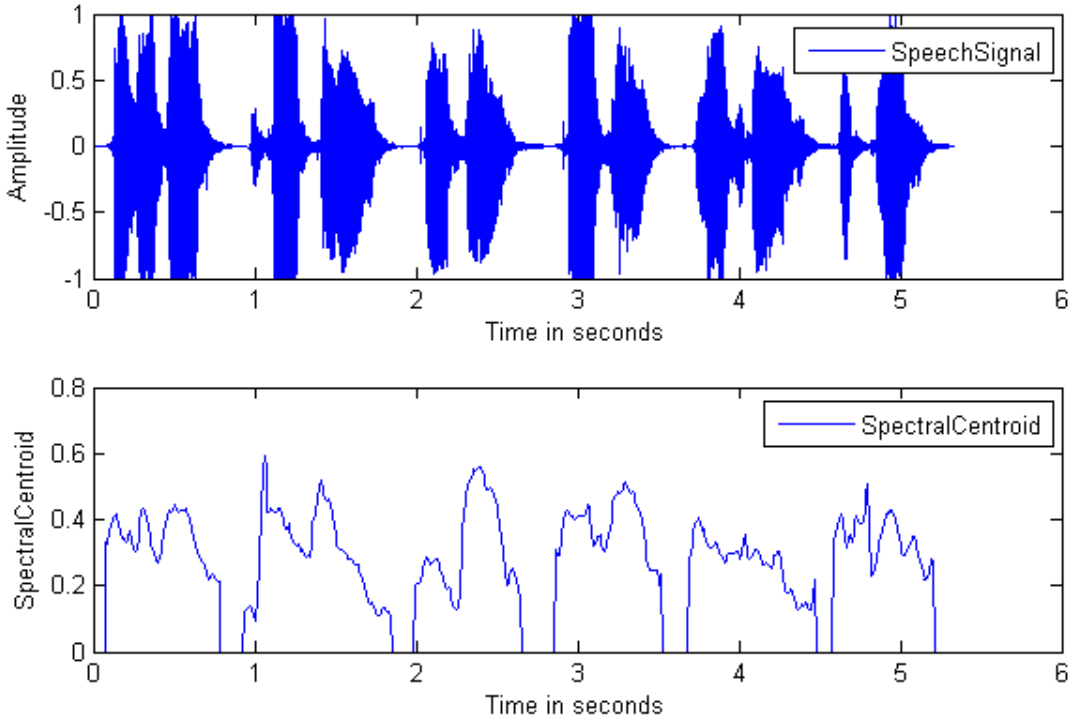
کمک می‌کنند، بلکه Spectral Centroid را به صورت دقیق‌تری بازتاب می‌دهند و دقت تحلیل را در کاربردهای مختلف افزایش می‌دهند.

Spectral Centroid همچنین در سیستم‌های تعاملی انسان-ماشین، نظیر دستیارهای صوتی، کاربرد گسترده‌ای دارد. این ویژگی به سیستم‌ها کمک می‌کند تا تمایز میان دستورالعمل‌های گفتاری مختلف را با توجه به تمبروال و روشنایی صداها شناسایی کنند. در ترکیب با سایر ویژگی‌های صوتی مانند MFCC و Spectral Centroid،Spectral Flux به‌طور موثری در توسعه سیستم‌های هوشمند و پیشرفته صوتی نقش ایفا می‌کند. به‌ویژه در سال‌های اخیر، استفاده از یادگیری عمیق برای تحلیل و ادغام Spectral Centroid با سایر ویژگی‌ها منجر به بهبود چشمگیر در دقت سیستم‌های پردازش صوت شده است. این پیشرفت‌ها به‌ویژه در سیستم‌های تشخیص احساسات و طبقه‌بندی صوتی مشاهده می‌شود.

نکته دیگری که Spectral Centroid را منحصربه‌فرد می‌کند، سادگی محاسبات آن است. این ویژگی می‌تواند با سرعت و کارآمدی بالایی از سیگنال‌های صوتی استخراج شود، که این موضوع برای سیستم‌های بلادرنگ نظیر تشخیص گفتار بسیار حیاتی است. همچنین، این ویژگی به دلیل همبستگی قوی با نحوه درک شنوایی انسان از روشنایی صدا، اغلب به عنوان معیاری مستقیم برای تحلیل تمبروال به کار گرفته می‌شود. با این حال، باید توجه داشت که Spectral Centroid به تنهایی نمی‌تواند اطلاعات جامعی از سیگنال صوتی ارائه دهد. ترکیب این ویژگی با ویژگی‌های دیگر نظیر Spectral Bandwidth و Zero Crossing Rate می‌تواند تحلیل جامع‌تری از سیگنال ارائه کند و دقت مدل‌های یادگیری ماشین را افزایش دهد.

در آینده، انتظار می‌رود که Spectral Centroid همچنان نقش برجسته‌ای در پیشرفت فناوری‌های پردازش صوت داشته باشد. با توسعه الگوریتم‌های پیشرفته یادگیری عمیق و پردازش موازی، قابلیت‌های این ویژگی در تحلیل دقیق‌تر صدا و بهبود سیستم‌های هوشمند بیشتر خواهد شد. این روند، به‌ویژه در زمینه‌هایی نظیر تعامل صوتی انسان-ماشین و تحلیل صدا در محیط‌های پیچیده و نویزی، تاثیر قابل‌توجهی خواهد داشت.

در تصویر زیر ، نمایش Spectral Centroid یک سیگنال صوتی نمایش داده شده است .



شکل ۹: Spectral Centroid یک سیگنال صوتی

۴_۵ ویژگی کروماتیک (chroma feature)

ویژگی کروماتیک، کیفیتی از یک کلاس زیر و بمی است که به "رنگ" یک زیر و بمی موسیقی اشاره دارد. این ویژگی می‌تواند به دو مقدار تجزیه شود: یک مقدار ثابت نسبت به اکتاو که "کروماتیک" نامیده میشود و یک "ارتفاع زیر و بمی" که نشاندهنده اکتاو زیر و بمی است.

در نُت‌نویسی موسیقی غربی، ۱۲ نت (کلاس زیر و بمی) وجود دارد که بر اساس فرکانس آن‌ها از C تا B در مجموعه $\{C, C\sharp, D, D\sharp, E, F, F\sharp, G, G\sharp, A, A\sharp, B\}$ مرتب شده‌اند که به آن ها کروماتیک گفته می‌شود. نتهای موجود در هر کلاس میتوانند در اکتاو از هم متفاوت باشند. در موسیقی هر اکتاو به معنای دو برابر(یا نصف) شدن فرکانس نت می‌باشد و به صورت یک عدد صحیح در کنار نت نشان داده میشود. به عبارتی اکتاوها به ما کمک میکنند تا نتها را در محدوده‌های فرکانسی مختلف دسته‌بندی کنیم. هر اکتاو شامل ۱۲ نت است که به ترتیب از C تا B مرتب شده‌اند و فرکانس هر نت در آن $\sqrt[12]{2}$ برابر نت قبلی است.

برای مثال فرکانس نت A3 و A#3 به ترتیب برابر ۲۲۰ و ۲۳۳.۰۸ هرتز و فرکانس نت A4 (دو برابر نت A3) ۴۴۰ هرتز می‌باشد. نت های A3 و A4 با اینکه در مقدار فرکانس تفاوت زیادی دارند ولی هر دو به کلاس A تعلق دارند. این کلاس‌بندی نت‌ها متناسب با ادراک شنوایی انسان است. زیرا اگرچه فرکانس این نتها متفاوت است، اما گوش انسان آنها را به عنوان نتهای مشابه در اکتاوهای مختلف درک میکند.

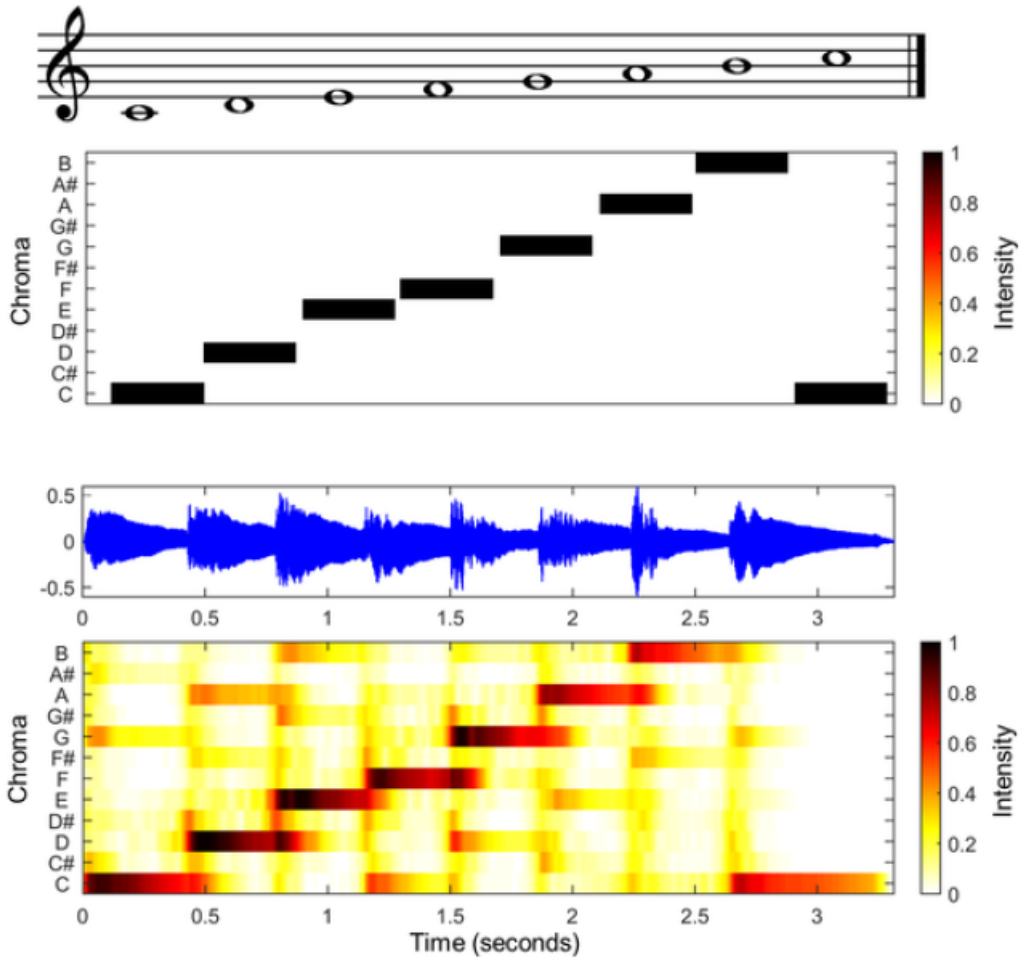
زیر و بمی تنها در صداهایی قابل تعیین است که فرکانس آنها به اندازه کافی واضح و پایدار باشد تا از نویز متمایز شود. برای استخراج ویژگی کروماتیک صدا باید ابتدا بازه هایی از فریم صدا را در نظر بگیریم و از تبدیل فوریه کوتاه‌مدت (STFT) مربوط به آن را محاسبه کنیم. این تبدیل فوریه شامل مولفه های فرکانسی مختلف است.

نزدیک ترین کلاس زیر و بمی مربوط به فرکانس این مولفه‌های فرکانسی صدا را بدست می‌آوریم. برای مثال ممکن است سه مولفه ی فرکانسی با فرکانس ۲۲۰ و دامنه ی ۱۰، فرکانس ۲۳۳ و دامنه ی ۸ و فرکانس ۴۴۱ با دامنه ی ۵ داشته باشیم. با توجه به فرکانس نت‌ها در اکتاوهای مختلف میتوان گفت که این مولفه‌های فرکانسی به ترتیب مربوط به نت‌های A3، A#3 و A4 هستند.

سپس باید مجموع انرژی و یا توان هر یک از کلاس‌های زیر و بمی بدست امده را با جمع کردن مقادیر دامنه‌ی آن ها (یا مجذور مقادیر دامنه) محاسبه کرد. برای این مثال می توان گفت که برای کلاس A خواهیم داشت ۱۵ و برای کلاس A# این مقدار ۸ خواهد بود.

مقادیر انرژی محاسبه شده برای هر یک از دوازده نت کروماتیک در یک وکتور ۱۲ عنصری قرار میگیرند. به این ترتیب وکتور کروماتیک مربوط به این زیربازه از صدا را می توان به صورت [0, 0, 0, 0, 0, 0, 0, 0, 0, 15, 8, 0] که مقادیر مربوط به هر کلاس زیر و بمی را نشان می‌دهد. در نهایت نیز این بردار نرمالیزه خواهد شد تا اندازه‌ی بردار یک شود.

به طور خلاصه، ویژگی کروماتیک، توزیع انرژی فرکانسی سیگنال صوتی را در دوازده کلاس زیر و بمی نشان میدهد. شکل زیر یک مثال از طیف مربوط به ویژگی کروماتیک بخشی از موسیقی در ۸ زیربازه زمانی را نشان می‌دهد.



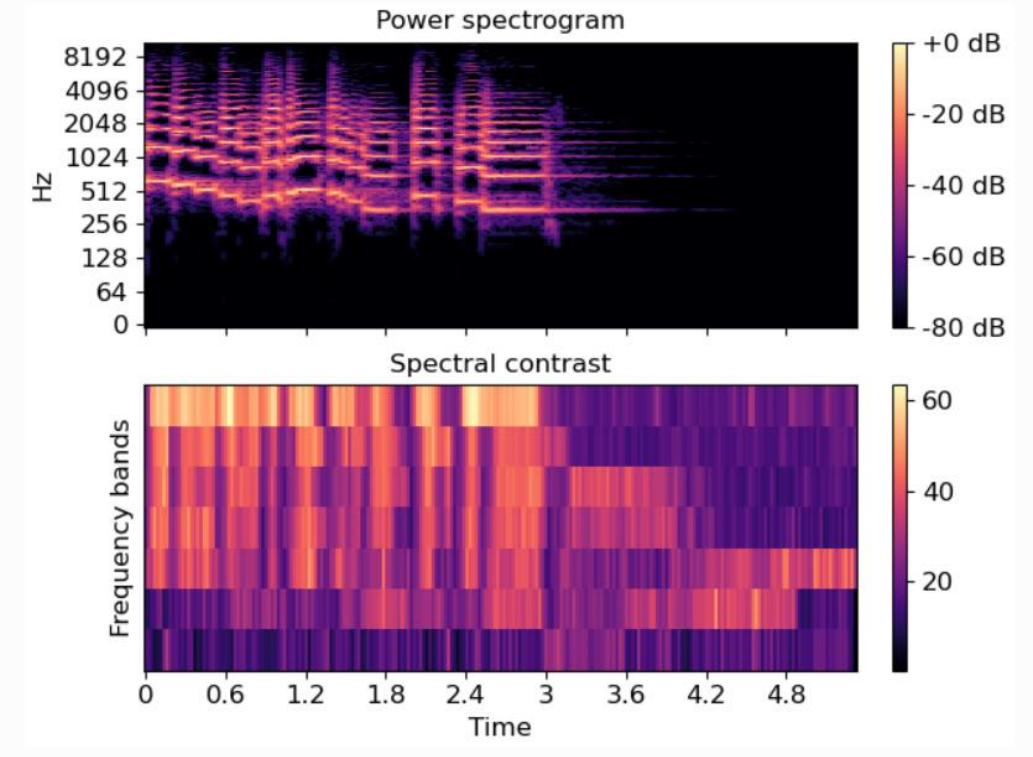
شکل ۱۰: (الف) موسیقی در مقیاس سی ماژور.

(ب) کروماگرام به دست آمده به صورت تئوری. (ج) صدای ضبط شده در مقیاس سی ماژور که روی پیانو پخش می شود.
(د)کروماگرام به دست آمده از ضبط صدا.

Spectoral Contrast ۶_۴

در تحلیل طیفی، هر فریم زمانی از یک اسپکتروگرام (طیف‌نگار) به زیر باندهای فرکانسی تقسیم میشود. برای هر زیر باند، کنتراست انرژی با مقایسه انرژی قله(میانگین انرژی در بالاترین صدک) با انرژی دره(میانگین انرژی در پایین‌ترین صدک) تخمین زده می‌شود. این روش کمک میکند تا تفاوتهای انرژی در هر زیر باند فرکانسی را در یک لحظه زمانی خاص بررسی کرد.

مقادیر بالای کنتراست معمولاً به سیگنالهای باریکباند و واضح اشاره دارند مثل یک نت موسیقی خاص، در حالی که مقادیر پایین کنتراست به نویز پهنباند(مانند صدای پس‌زمینه) اشاره میکنند. این روش به شناسایی و تمایز بین بخشهای مختلف سیگنال کمک میکند و میتواند در تحلیل موسیقی و سایر کاربردهای صوتی مفید باشد.



شکل ۱۱: طیف‌نگار مربوط به طیف تضاد

۷_۴ (ZCR)Zero-Crossing Rate

یک ویژگی پراستفاده در تحلیل صدا و موسیقی است که نشان‌دهنده تعداد دفعات عبور سیگنال صوتی از محور صفر در یک فریم صوتی است. ZCR نشان می‌دهد که سیگنال صوتی چند بار در یک فریم زمانی خاص، علامت خود را از مثبت به منفی یا برعکس تغییر داده است. این ویژگی اغلب برای تحلیل سیگنال‌های غیراستاندار (مانند سیگنال‌های صوتی) استفاده می‌شود. ZCR با شمارش تعداد گذر از محور ۰(تعداد تغییرات علامت سیگنال) در فریم‌های تقسیم‌شده صوت محاسبه می‌شود. ابزارهایی مانند Librosa در پایتون معمولاً برای استخراج سریع این ویژگی استفاده می‌شوند.

۴_۷_۱ کاربردهای ZCR

یکی از کاربردهای اصلی ZCR در تفکیک بخش‌های صدادار (voiced) و بی‌صدا (unvoiced) است. صداهای صدادار (مانند مصوت‌ها) معمولاً ZCR کمتری دارند، در حالی که صداهای بی‌صدا (مانند اصواتی مثل "س" یا "ش") ZCR بیشتری نشان می‌دهند. در موسیقی، ZCR به شناسایی صداهای ضربی با شروع‌های تیز و نرخ بالای عبور از صفر کمک می‌کند. علاوه‌بر این، می‌تواند در تفکیک نویز از سیگنال‌های مفید کاربرد داشته باشد؛ زیرا نویز معمولاً ZCR بیشتری نسبت به گفتار یا موسیقی دارد. در وظایف طبقه‌بندی مانند تشخیص صدای مردانه و زنانه کاربرد دارد. صداهای مردانه معمولاً به دلیل زیر بودن فرکانس‌ها ZCR کمتری دارند.

۴_۷_۲ چالش‌های استفاده از ZCR

سیگنال‌هایی با دامنه کم یا بخش‌های تقریباً بی‌صدا می‌توانند منجر به نتایج غلط در ZCR شوند. تکنیک‌هایی مانند افزودن یک مقدار ثابت کوچک یا تنظیم شرایط پیش‌پردازش برای کاهش این خطا استفاده می‌شوند. برای پایدارسازی اندازه‌گیریZCR، از تکنیک‌هایی مثل پنجره‌بندی استفاده می‌شود.

ZCR یک ویژگی ساده ولی موثر در پردازش سیگنال‌های صوتی است که طیف وسیعی از کاربردها در تحلیل گفتار، موسیقی و حذف نویز دارد. پیشرفت‌های موجود در ابزارهای محاسباتی و ترکیب ZCR با ویژگی‌های طیفی، کاربردهای آن را بیشتر گسترش داده است.

۴_۸ Linear Predictive Coding (LPC)

یک تکنیک قوی استخراج ویژگی است که در پردازش سیگنالهای صوتی استفاده می‌شود، و به طور خاص در شناسایی صدا و آنالیز گفتار کاربرد دارد. این روش با مدل‌سازی دستگاه تولید صدا در انسان به عنوان یک سیستم خطی و تقریب سیگنال صوتی از طریق پیش‌بینی خطی عمل می‌کند.

نمایش سیگنال LPC سیگنال گفتار را به‌صورت ترکیبی از نمونه‌های قبلی و یک مقدار خطا مدل می‌کند. ضرایب این مدل برای کاهش خطای پیش‌بینی استخراج می‌شوند. فرمانت‌ها که فرکانس‌های تشدید دستگاه صوتی هستند، توسط LPC شناسایی می‌شوند و برای تمایز بین صداهای مختلف اهمیت زیادی دارند.

LPC سیگنال صوتی را فشرده‌سازی می‌کند و به‌جای شکل موج خام، ضرایب پیش‌بینی خطی را رمزگذاری می‌کند. این ویژگی در کدک‌های GSM بسیار استفاده می‌شود.

برای پیاده‌سازی LPC، سیگنال گفتار به فریم‌های کوتاه، معمولاً ۱۰ تا ۳۰ میلی‌ثانیه، تقسیم می‌شود. سپس، فیلترهایی برای تقویت فرکانس‌های بالاتر به سیگنال اعمال می‌شوند تا پاسخ فرکانسی هموارتر شود. در نهایت، الگوریتم‌هایی مانند *Levinson-Durbin* ضرایبی را محاسبه می‌کنند که سیگنال صوتی را به‌خوبی مدل‌سازی کنند.

۴_۸_۱ کاربردهای LPC

ویژگی‌های استخراج‌شده توسط LPC برای ساخت مدل‌های صوتی خاص هر گوینده استفاده می‌شود که در سیستم‌های احراز هویت زیستی کاربرد دارند. علاوه‌بر این، برای فشرده‌سازی و بازسازی گفتار در ارتباطات موبایل مانند (GSM) استفاده می‌شود که انتقال داده کارآمدتری را ممکن می‌سازد.

تلفیق گفتار LPC می‌تواند با تحریک یک فیلتر دیجیتالی توسط اصوات صدادار یا بی‌صدا، گفتار مصنوعی تولید کند. در تحلیل و شناسایی احساسات موجود در گفتار ترکیب LPC با یادگیری ماشین مؤثر است.

پژوهش‌ها بر ترکیب ویژگی‌های LPC با مدل‌های یادگیری عمیق تمرکز دارند تا دقت در محیط‌های پر نویز را افزایش دهند، تحلیل لهجه‌های متنوع را بهبود بخشند، و توانایی شناسایی احساس را بهبود دهند. همچنین، LPCبه‌صورت ترکیبی با تکنیک‌های پیشرفته‌ای مانند ضرایب طیفی فرکانس مل (MFCC) استفاده می‌شود تا کاربردهای آن گسترده‌تر شود.

۴_۹ Perceptual Linear Prediction (PLP)

یکی دیگر از تکنیک‌های پرکاربرد در پردازش گفتار و صدا است که با استفاده از اصول روان‌صوت‌شناسی (پرسپتوال)، ویژگی‌های آکوستیکی سیگنال را بهبود می‌دهد. این تکنیک به‌طور گسترده‌ای برای تشخیص خودکار گفتار و کاربردهای مشابه استفاده می‌شود.

مدل‌سازی PLP از مدل‌هایی بهره می‌برد که بر اساس نحوه درک انسان از صدا طراحی شده‌اند، از جمله:

- ادغام باند بحرانی که حساسیت گوش به فرکانس‌های مختلف را شبیه‌سازی می‌کند.
- وزن‌دهی بلندی برابر که نشان‌دهنده حساسیت متغیر گوش به فرکانس‌های مختلف است.
- فشرده‌سازی شدت به بلندی برای تطبیق با ادراک غیرخطی انسان از شدت صدا.

استخراج ویژگی با استفاده از روش PLP شامل چندین مرحله است: ابتدا صدای ورودی فیلتر می‌شود تا نویزهای اضافی حذف شوند. سپس طیف فرکانسی صدا با استفاده از تبدیل فوریه به‌دست می‌آید. بعد از آن، طیف فرکانسی به باندهای فرکانسی کوچکتر تقسیم و در هر باند انرژی محاسبه می‌شود. این مقادیر انرژی به واحدهای دسی‌بل تبدیل می‌شوند و یک تبدیل غیرخطی برای شبیه‌سازی پاسخ گوش انسان به فرکانس‌ها انجام می‌گیرد. در نهایت، با استفاده از پیش‌بینی خطی(LPC)، ویژگی‌های نهایی استخراج می‌شوند که برای تشخیص و طبقه‌بندی صداها به کار می‌روند.

۴_۹_۱ کاربردهای PLP

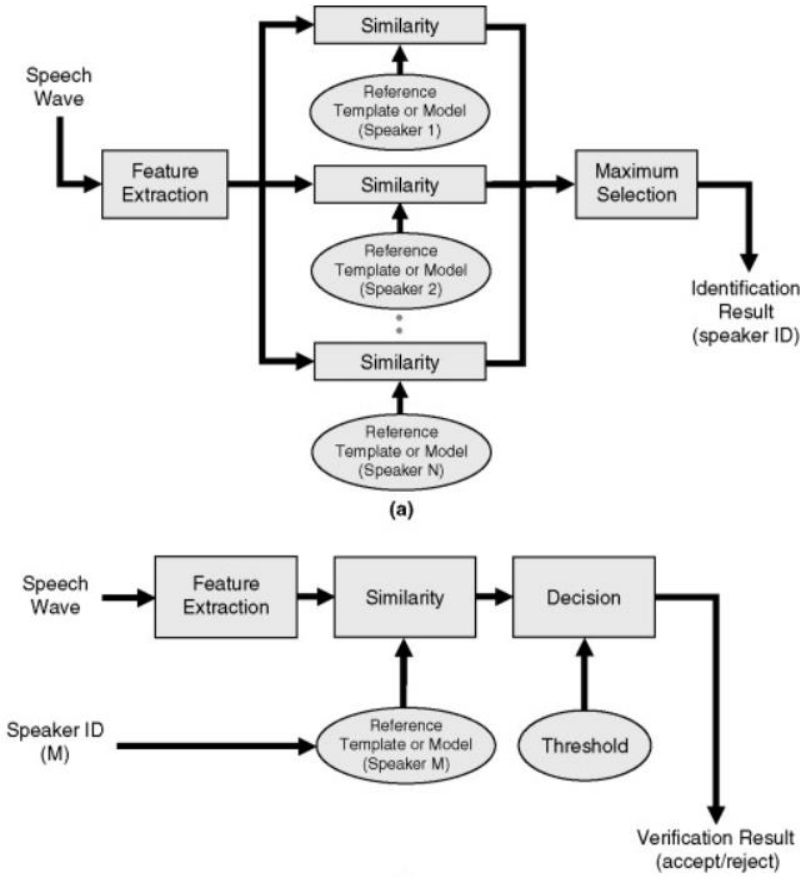
PLP با تمرکز بر ویژگی‌های مهم ادراکی، برای شناسایی گفتار حتی در محیط‌های نویزی مناسب است. این روش می‌تواند ویژگی‌های مستقل از گوینده را جدا کند در حالی که ویژگی‌های مخصوص هر گوینده را برای شناسایی حفظ می‌کند.

PLP یکی از تکنیک‌های کلیدی در پردازش گفتار است که ویژگی‌های آن در ترکیب با یادگیری ماشین برای بهبود عملکرد سیستم‌ها به‌ویژه در شرایط پر نویز به‌کار می‌رود. PLP با توجه به ادراک انسانی از صدا، نمایش دقیق‌تری ارائه می‌دهد و برای سیستم‌های خودکار کارایی بیشتری دارد.

۵ یادگیری شباهت (similarity learning)

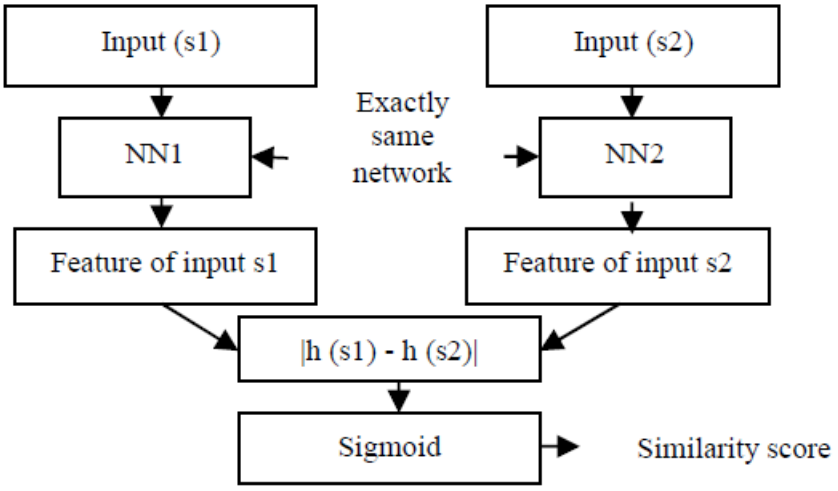
یادگیری شباهت در تجزیه و تحلیل صوتی، شامل یادگیری تابعی برای اندازه‌گیری شباهت یا عدم شباهت دو سیگنال صوتی است. این نوع یادگیری یکی از انواع یادگیری نظارتی محسوب می‌شود که شامل ایجاد نمایش برداری از نمونه‌های صوتی و سپس مقایسه‌ی آن‌ها برای اندازه‌گیری شباهت است. این مدل یادگیری در کاربردهایی مانند شناسایی و احراز هویت گوینده بر اساس سیگنال صوتی استفاده می‌شود.

شکل زیر ساختار پایه‌ی شناسایی و احراز هویت گوینده را بر مبنای محاسبه‌ی شباهت بین سیگنال‌های صوتی گویندگان مختلف نشان می‌دهد. در شناسایی گوینده، سیستم تلاش میکند از بین گویندگان موجود، گوینده‌ی گفتار ورودی را مشخص کند. اما در احراز هویت گوینده باید مشخص کند که آیا هویت ادعا شده توسط گوینده پذیرفته یا رد می‌شود.



شکل ۱۲: ساختار پایه شناسایی و احراز هویت گوینده

یادگیری شباهت ارتباط بسیار نزدیکی با مفاهیم رگرسیون و طبقه‌بندی دارد. شبکه های Triple Network و Siames Network از جمله مدل‌های یادگیری شباهت مبتنی بر یادگیری عمیق هستند. شکل زیر معماری کلی شبکه‌ی Siames را نشان می‌دهد. این شبکه در کاربردهای تحلیل گفتار از دو شبکه عصبی CNN یکسان به صورت موازی تشکیل شده است. این شبکه‌های CNN، اکسپکتوگرام (طیف مربوط به صوت) را به عنوان ورودی دریافت می‌کند. در شبکه‌های Siames دو ورودی به صورت همزمان پردازش شده و خروجی آن نیز میزان شباهت دو صوت ورودی می‌باشد. این شبکه نیز به گونه‌ای آموزش میبیند که فاصله‌ی نمونه‌های مشابه به هم نزدیک و نمونه‌های غیرمشابه از هم دور باشد. این مدل‌های یادگیری از تابع هزینه‌ی Contrastive loss استفاده می‌کند.



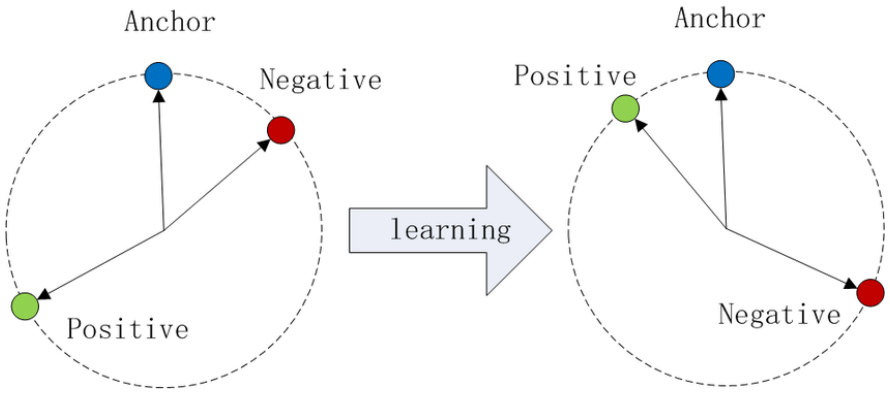
شکل ۱۳: معماری پایه شبکه siames

در شبکه‌های Triple برای آموزش نیاز به سه نمونه با عنوان‌های anchor(صدای یک گوینده)، نمونه‌ی مثبت(صدای دیگر همان گوینده) و نمونه‌ی منفی(صدای گوینده‌ی دیگر) است. این شبکه به گونه‌ای آموزش میبیند که فاصله‌ی anchor با نمونه‌ی مثبت کم باشد و در عین حال از نمونه‌ی منفی دور باشد. در واقع این شبکه، سه نمونه را به صورت همزمان پردازش می‌کند و از تابع هزینه‌ی Triple loss برای بهبود دقت در تشخیص شباهت استفاده می‌کند.

۵_۱ توابع هزینه‌ی رایج در یادگیری شباهت:

در یادگیری شباهتی باید معیارهایی برای شباهت در نظر گرفته شود برای مثال میتوان فاصله‌ی اقلیدسی و یا فاصله‌ی cosine را در نظر گرفت. دو تابع هزینه‌ی رایج در این حوزه نیز Triple loss و Contrastive loss می‌باشد.

تابع هزینه Triple loss: در این تابع هزینه سه نوع نمونه به عنوان ورودی گرفته میشوند که به آنها نمونههای مثبت، نمونههای هدف(anchor) و نمونههای منفی گفته میشود. در این میان، دو نمونه اول صداهای مختلف از یک گوینده هستند و دو نمونه آخر صداهای گویندگان مختلف هستند. هدف از آموزش این است که شباهت (برای مثال شباهت کسینوسی) دو نمونه اول بیشتر از دو نمونه آخر باشد. تغییرات فاصله‌ی بین نمونه‌های ورودی در روند آموزش، برمبنای این تابع هزینه در شکل آورده شده است.



شکل ۱۴: تغییرات فاصله نمونه ها در روند آموزش بر اساس تابع هزینه Triple loss

تابع هزینه Contrastive loss: این تابع هزینه نیز مانند تابع هزینه‌ی Triple loss می‌باشد با این تفاوت که در آن ورودی‌ها به صورت دو نمونه‌ی ورودی شبیه یا دو نمونه‌ی ورودی متفاوت است و هدف از این تابع هزینه، کم کردن فاصله بین نمونه‌های مشابه و زیاد کردن فاصله بین نمونه‌های متفاوت در روند یادگیری می‌باشد.

هر دوی این توابع هزینه، نیاز به انتخاب دقیق نمونه‌ها برای آموزش مؤثر دارند. در تابع Triple loss، باید سه‌تایی‌هایی از نمونه‌ها انتخاب شوند که شامل یک نمونه مرجع، یک نمونه مثبت و یک نمونه منفی

باشند. در تابع Contrastive loss، باید جفت‌هایی از نمونه‌ها انتخاب شوند که مشابه یا متفاوت باشند. این فرآیند میتواند زمانبر و محاسباتی پرهزینه باشد و اگر نمونه‌ها به دقت انتخاب نشوند، ممکن است منجر به بیش‌برازش (overfitting) شود.

۶ مراجع

[1] N. Chandolikar, C. Joshi, P. Roy, A. Gawas, and M. Vishwakarma, “Voice recognition: A comprehensive survey,” in *2022 International Mobile and Embedded Technology Conference (MECON)*, 2022.

[2] N. H. Tandel, H. B. Prajapati, and V. K. Dabhi, “Voice recognition and voice comparison using machine learning techniques: A survey,” in *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, 2020.

[3] B. Yelure, S. Patil, A. Nayakwadi, C. Raut, K. Joshi, and A. Nadaf, “Machine Learning based Voice Authentication and Identification,” in *2023 3rd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, 2023, pp. 936–940.

[4] “Feature extraction — librosa 0.10.2.post1 documentation,” *Librosa.org*. [Online]. Available: https://librosa.org/doc/main/feature.html. [Accessed: 30-Dec-2024].

[5] *Kaggle.com*. [Online]. Available: https://www.kaggle.com/code/gopidurgaprasad/mfcc-feature-extraction-from-audio. [Accessed: 30-Dec-2024].

[6] M. S. Ahmad, “Deep learning 101: Lesson 23: The basics of audio signal processing with FFT,” *Medium*, 02-Sep-2024. [Online]. Available: https://muneebsa.medium.com/deep-learning-101-lesson-23-the-basics-of-audio-signal-processing-with-fft-ffef65689c1d. [Accessed: 30-Dec-2024].

[7] B. A. Alsaify, H. S. Abu Arja, B. Y. Maayah, M. M. Al-Taweel, R. Alazrai, and M. I. Daoud, “Voice-Based Human Identification using Machine Learning,” in *2022 13th International Conference on Information and Communication Systems (ICICS)*, 2022.

[8] R. Sharma, D. Govind, J. Mishra, A. K. Dubey, K. T. Deepak, and S. R. M. Prasanna, “Milestones in speaker recognition,” *Artif. Intell. Rev.*, vol. 57, no. 3, 2024.

[9] K. W. Cheuk, H. Anderson, K. Agres, and D. Herremans, “NnAudio: An on-the-fly GPU audio to spectrogram conversion toolbox using 1D convolutional neural networks,” *IEEE Access*, vol. 8, pp. 161981–162003, 2020.

[10] C. Li *et al.*, “Deep Speaker: An end-to-end neural speaker embedding system,” *arXiv [cs.CL]*, 2017.

[11] V. K. Pande, V. K. Kale, and S. Tharewal, “Audio data feature extraction for speaker diarization,” in *Proceedings of the NIELIT’s International Conference on Communication, Electronics and Digital Technology*, Singapore: Springer Nature Singapore, 2024, pp. 243–255.

[12] W. N. Jasim, S. A. W. Saddam, and E. J. Harfash, “Wind sounds classification using different audio feature extraction techniques,” *Informatica (Ljubl.)*, vol. 45, no. 7, 2022.

[13] S. Vijayputra, *Gender Recognition Using Fast Fourier Transform With Ann*. 2019.

[14] S. Furui, “Speaker Recognition in Smart Environments,” in *Human-Centric Interfaces for Ambient Intelligence*, Elsevier, 2010, pp. 163–184.