دانشگاه تهران

پردیس دانشکده های فنی

دانشکده برق و کامپیوتر

پروژه نهایی یادگیری ماشین

اعضای گروه به همراه شماره دانشجویی

مصطفی کرمانی‌نیا - 810101575

آیدین کاظمی - 810101561

امیر نداف فهمیده - 810101540

طاها مجلسی - 810101504

اساتید درس

دکتر اعرابی

دکتر ابوالقاسمی

بهمن 1403

فهرست مطالب

بخش اول : گزارش اولیه

بخش دوم: گزارش نهایی

در نوت بوک و به زبان انگلیسی موجود است. (لینک نوت بوک)

مقدمه ای بر voice authentication

تعریف voice authentication و اهمیت آن

احراز هویت صوتی (Voice Authentication) روشی برای شناسایی و تایید هویت افراد از طریق صدای آنهاست. صدای هر فرد به دلیل ویژگی‌های بیولوژیکی منحصربه‌فردش مانند شکل حنجره، اندازه دهان و الگوهای گفتاری، به‌عنوان یک مشخصه بیومتریک قابل‌اعتماد شناخته می‌شود. این تکنولوژی از این ویژگی‌ها برای ایجاد یک شناسه منحصربه‌فرد برای هر کاربر استفاده می‌کند.

مراحل کارکرد:

1. ایجاد پروفایل صوتی (Voice Enrollment): در مرحله اول، کاربر صدای خود را با گفتن عبارت یا کلمه مشخصی ضبط می‌کند. این صدا توسط سیستم پردازش می‌شود و ویژگی‌های کلیدی آن مانند فرکانس، تُن، ریتم و شدت استخراج و به‌عنوان یک الگوی مرجع ذخیره می‌شود.
2. تحلیل ویژگی‌های صوتی: سیستم با استفاده از تکنولوژی‌های پیشرفته پردازش سیگنال صوتی (Speech Signal Processing) و یادگیری ماشین، ویژگی‌های بیومتریک صدای کاربر را ثبت می‌کند.[18] این ویژگی‌ها شامل فرکانس‌های پایه صدا، زیر و بمی صدا، نحوه تاکید روی کلمات و الگوهای منحصر به‌فرد در تنفس و مکث‌های طبیعی می‌شوند.
3. فرایند شناسایی یا تایید هویت: هنگام احراز هویت، کاربر عبارت مشخصی را دوباره بیان می‌کند. صدای جدید با الگوی مرجع ذخیره‌شده مقایسه می‌شود. اگر تطابق کافی وجود داشته باشد، سیستم هویت را تأیید می‌کند.

در دنیای امروز که امنیت اطلاعات و دسترسی آسان به خدمات دیجیتال اهمیت بسیاری دارد، احراز هویت صوتی به‌عنوان یکی از نوآورانه‌ترین و مؤثرترین روش‌های شناسایی کاربران مطرح شده است. این فناوری نه‌تنها سطح امنیت را بهبود می‌بخشد، بلکه تجربه‌ای سریع و ساده را برای کاربران فراهم می‌کند برخی از اهمیت های آن به شرح زیر است.

1. امنیت بالا و حفاظت از حریم خصوصی:

هر فرد صدایی منحصربه‌فرد دارد که ناشی از ساختار بیولوژیکی و الگوهای گفتاری اوست. این ویژگی، جعل یا سوءاستفاده از صدا را بسیار دشوار می‌کند و امنیت سیستم‌ها را تضمین می‌کند.

2. سهولت استفاده:

در مقایسه با رمزهای عبور پیچیده یا کارت‌های شناسایی، احراز هویت صوتی نیازی به ابزارهای فیزیکی یا به‌خاطر سپردن اطلاعات ندارد. کاربران تنها با گفتن یک عبارت کوتاه می‌توانند به اطلاعات و خدمات خود دسترسی پیدا کنند.

3. کاهش هزینه‌ها و زمان:

این روش به دلیل عدم نیاز به سخت‌افزارهای گران‌قیمت یا زیرساخت‌های پیچیده، راهکاری مقرون‌به‌صرفه است و زمان لازم برای تایید هویت را به حداقل می‌رساند.

4. کاربرد گسترده در صنایع مختلف:

از بانکداری و خدمات مالی گرفته تا دستگاه‌های هوشمند و سیستم‌های امنیتی، احراز هویت صوتی به‌عنوان یک ابزار چندمنظوره می‌تواند در بسیاری از حوزه‌ها به‌کار گرفته شود و تجربه‌ای کاربر محور ایجاد کند.

5. تقویت تجربه مشتری:

در دنیای رقابتی امروز، ارائه تجربه‌ای روان و بدون دردسر به مشتریان یک مزیت کلیدی است. احراز هویت صوتی این امکان را فراهم می‌کند که کاربران در کمترین زمان و بدون مراحل پیچیده به خدمات مورد نظرشان دسترسی پیدا کنند.

بررسی کاربرد های voice authentication مانند شناسایی گوینده و تشخیص جنسیت گوینده

احراز هویت صوتی (Voice Authentication) یکی از پیشرفته‌ترین فناوری‌های شناسایی بیومتریک است که علاوه بر تأیید هویت، می‌تواند اطلاعات ارزشمند دیگری مانند جنسیت، احساسات و حتی زبان و لهجه گوینده را استخراج کند. کاربرد های هر یک از این اطلاعات به همراه توضیح مختصر به شرح زیر است.

1. شناسایی گوینده (Speaker Identification)

شناسایی گوینده فرآیندی است که در آن، سیستم تلاش می‌کند صدای یک فرد خاص را از میان گروهی از افراد تشخیص دهد. این کار با تطبیق الگوهای صوتی ضبط شده با صدای ورودی انجام می‌شود.

کاربردها:

- امنیت سازمانی: شناسایی و تایید هویت کارکنان مجاز برای دسترسی به اطلاعات حساس یا محیط‌های خاص.
- تحقیقات قضایی: شناسایی هویت افراد از طریق صدای ضبط‌شده در مکالمات یا پیام‌های صوتی.
- خدمات دیجیتال: ورود به حساب‌های کاربری یا تراکنش‌های مالی تنها با صدای کاربر.
- شخصی‌سازی خدمات: مانند ایجاد پروفایل‌های اختصاصی برای کاربران در دستیارهای صوتی مانند Alexa یا Siri.

2. تشخیص جنسیت گوینده (Gender Recognition)

تشخیص جنسیت گوینده یکی دیگر از قابلیت‌های احراز هویت صوتی است که با تحلیل فرکانس‌های صوتی و نحوه بیان، جنسیت فرد را مشخص می‌کند. صدای مردان به دلیل فرکانس پایین‌تر و صدای زنان به دلیل فرکانس بالاتر از یکدیگر متمایز می‌شود.

کاربردها:

- دستیارهای صوتی هوشمند: تنظیم پاسخ‌ها بر اساس جنسیت کاربر برای ایجاد تعامل طبیعی‌تر.
- تبلیغات دیجیتال: ارائه محتواهای متناسب با جنسیت شناسایی‌شده برای افزایش اثربخشی تبلیغات.

3. تحلیل احساسات (Emotion Recognition)

تحلیل احساسات یکی از قابلیت‌های مهم فناوری احراز هویت صوتی است که از طریق تغییرات در تُن صدا، شدت و ریتم گفتار می‌تواند حالات احساسی فرد را شناسایی کند. این تکنولوژی می‌تواند احساساتی مانند شادی، خشم، اضطراب یا ناراحتی را تشخیص دهد.

کاربردها:

- سلامت روانی: در برنامه‌های نظارتی یا درمانی، تحلیل احساسات می‌تواند به شناسایی مشکلات روحی و روانی کاربران کمک کند.
- تعاملات انسانی هوشمند: در ربات‌های اجتماعی و دستیارهای صوتی، تحلیل احساسات برای تنظیم رفتار و لحن پاسخ‌دهی به کار می‌رود.
- بهبود تجربه مشتری: در سیستم‌های پشتیبانی، شناسایی احساسات مشتری می‌تواند به ارائه پاسخ‌های مناسب‌تر و رفع سریع‌تر مشکلات کمک کند.

4. تشخیص زبان و لهجه (Language and Accent Recognition)

تشخیص زبان و لهجه به سیستم اجازه می‌دهد زبان گوینده یا ویژگی‌های لهجه خاص او را شناسایی کند. این قابلیت از تحلیل ویژگی‌های صوتی، مانند نحوه بیان و استرس در کلمات، برای دسته‌بندی زبان و لهجه استفاده می‌کند.

کاربردها:

- سیستم‌های ترجمه: شناسایی زبان گوینده برای ارائه ترجمه‌های دقیق‌تر.
- آموزش زبان: ایجاد برنامه‌های آموزشی بر اساس زبان مادری یا لهجه کاربران.
- شخصی‌سازی محتوا: ارائه محتوا و تبلیغات متناسب با زبان یا لهجه گوینده.

5. اینترنت اشیا (IoT) و کنترل هوشمند

احراز هویت صوتی به کاربران اجازه می‌دهد تنها با صدای خود دستگاه‌های هوشمند از تنظیم دمای خانه گرفته تا دسترسی به اطلاعات حساس در محیط‌های صنعتی را کنترل کنند.

کاربردها:

- قفل‌گشایی در خانه‌های هوشمند.
- کنترل دستگاه‌های صنعتی توسط اپراتورهای تأییدشده.
- دسترسی ایمن به خودروها یا دستگاه‌های همراه.

تعریف closed-set authentication و open-set authentication و تفاوت های این دو با هم

تعریف closed-set authentication

احراز هویت مجموعه بسته (Closed-Set Authentication) به سیستمی اطلاق می‌شود که در آن تنها کاربران از پیش تعریف‌شده و شناخته‌شده مجاز به دسترسی هستند.

ویژگی‌های کلیدی closed-set authentication:

- محدودیت به کاربران مشخص: فقط افرادی که قبلاً در سیستم ثبت شده‌اند، می‌توانند احراز هویت شوند.
- سرعت و دقت بالا: به دلیل محدود بودن مجموعه کاربران، فرآیند تطبیق سریع‌تر و با دقت بیشتری انجام می‌شود.

- کاربرد در محیط‌های کنترل‌شده: این روش برای سازمان‌ها و محیط‌هایی مناسب است که نیاز به کنترل دقیق دسترسی دارند.

کاربردها:

- سیستم‌های بیومتریک: در سیستم‌های تشخیص هویت بر اساس ویژگی‌های بیومتریک مانند اثر انگشت یا چهره، اغلب از رویکرد مجموعه بسته استفاده می‌شود تا فقط افراد مجاز شناسایی شوند.
- سیستم‌های امنیتی سازمانی: در محیط‌های کاری که نیاز به دسترسی محدود به منابع وجود دارد، این روش می‌تواند اطمینان حاصل کند که فقط کارکنان مجاز به اطلاعات حساس دسترسی دارند.

محدودیت‌ها:

- عدم انعطاف‌پذیری: در مواجهه با کاربران جدید یا تغییرات در مجموعه کاربران، نیاز به به‌روزرسانی مداوم سیستم وجود دارد.
- عدم توانایی در شناسایی تهدیدات جدید: کاربران مخرب جدید که در سیستم ثبت نشده‌اند، ممکن است شناسایی نشوند.

تعریف open-set authentication

احراز هویت مجموعه باز یکی از روش‌های پیشرفته امنیتی است که هدف آن شناسایی و مدیریت هم کاربران شناخته‌شده و هم کاربران ناشناخته‌ای است که به سیستم دسترسی پیدا می‌کنند. برخلاف احراز هویت مجموعه بسته (Closed-Set Authentication) که تنها کاربران ثبت‌شده را شناسایی می‌کند، در این روش سیستم توانایی تشخیص و برخورد با کاربران یا دستگاه‌های جدید و ناشناخته را دارد.

ویژگی‌های کلیدی open-set authentication:

- تشخیص کاربران ناشناخته: این سیستم می‌تواند افرادی را که در پایگاه داده ثبت نشده‌اند، شناسایی کرده و به‌عنوان "ناشناخته" علامت‌گذاری کند.
- افزایش امنیت: با شناسایی کاربران جدید یا ناشناخته، احتمال نفوذ غیرمجاز کاهش پیدا می‌کند. به طور مثال سیستم قادر است کاربرانی که ناشناخته هستند را شناسایی کرده و دسترسی آن‌ها را محدود کند و به کمک این توانایی مانع از نفوذ هکرها یا کاربران غیرمجاز به سیستم می‌شود.
- انعطاف‌پذیری: این روش برای محیط‌هایی که کاربران یا دستگاه‌های جدید به‌طور مداوم اضافه می‌شوند، مانند شبکه‌های IoT، بسیار مفید است.

کاربردها:

- اینترنت اشیا (IoT): در شبکه‌های IoT، دستگاه‌های جدید مرتباً به سیستم اضافه می‌شوند. احراز هویت مجموعه باز کمک می‌کند که دستگاه‌های غیرمجاز شناسایی شوند و امنیت شبکه حفظ شود.
- ارتباطات بی‌سیم: در سیستم‌هایی مانند RFF(Radio Frequency Fingerprinting)، از روش مجموعه باز برای شناسایی ارسال‌کنندگان شناخته‌شده و رد دستگاه‌های ناشناخته استفاده می‌شود.
- مدل‌های هوش مصنوعی و یادگیری ماشین: احراز هویت مجموعه باز به مدل‌های یادگیری ماشین کمک می‌کند تا ورودی‌های ناشناخته را شناسایی کنند و از تصمیم‌گیری اشتباه جلوگیری شود.

اما استفاده از open-set authentication چالش‌های مخصوص به خود دارد که عبارتند از:

- عمومیت مدل‌ها: طراحی مدل‌هایی که بتوانند بدون داشتن اطلاعات قبلی، کاربران یا داده‌های ناشناخته را شناسایی کنند، پیچیده است.
- کمبود داده‌های ناشناخته: نبود داده‌های کافی برای آموزش مدل‌ها در مواجهه با تهدیدات ناشناخته، دقت سیستم را کاهش می‌دهد.

- پیچیدگی محاسباتی: شناسایی کاربران ناشناخته می‌تواند به قدرت پردازشی بیشتری نیاز داشته باشد که برای سیستم‌های محدود مانند IoT چالش‌برانگیز است.

تفاوت های این دو (به صورت جمع بندی)

با خواندن توضیحات هر بخش به طور کلی متوجه تفاوت این دو روش شدیم اما در این بخش به صورت جمع بندی کاربردها، مزایا، معایب، تکنیک های استفاده شده و چالش‌های هر کدام را بیان می‌کنیم و در آخر به یک جمع بندی می‌رسیم.

مقایسه کاربردها: با توجه به تعاریف هر روش می‌توان گفت که Closed-Set Authentication مناسب برای محیط‌هایی با تعداد کاربران ثابت و محدود، مثل سیستم‌های داخلی شرکت‌ها، سیستم‌های کنترل دسترسی فیزیکی و سیستم‌های کوچک بیومتریک است در حالی که Open-Set Authentication مناسب برای محیط‌های پویا و غیرقابل پیش‌بینی که کاربران یا دستگاه‌های جدید به‌طور مداوم اضافه می‌شوند، مانند شبکه‌های IoT، سرویس‌های آنلاین و سیستم‌های بیومتریک عمومی است.

بررسی مزایای هر روش:

:Closed-Set Authentication

- پیاده‌سازی ساده: سیستم فقط به داده‌های کاربران شناخته‌شده نیاز دارد.
- دقت بالا در شناسایی کاربران ثبت‌شده: با داده‌های محدود و دقیق‌تر، نرخ خطا کاهش می‌یابد.
- کارایی بالا در محیط‌های محدود: برای سازمان‌هایی که کاربران آن‌ها از پیش تعیین‌شده هستند، عملکردی مطلوب دارد.

:Open-Set Authentication

- شناسایی کاربران ناشناخته: توانایی مدیریت کاربران جدید یا تهدیدات غیرمنتظره.

- امنیت پیشرفته‌تر: سیستم می‌تواند به تهدیدات جدید واکنش نشان دهد و نفوذ غیرمجاز را کاهش دهد.
- انعطاف‌پذیری بالا: مناسب برای محیط‌های پویا با کاربران متغیر.

بررسی معایب هر روش:

:Closed-Set Authentication

- عدم توانایی در شناسایی تهدیدات جدید: کاربران غیرمجاز که در سیستم ثبت نشده‌اند شناسایی نمی‌شوند.
- نیاز به به‌روزرسانی مداوم: با اضافه شدن کاربران جدید، پایگاه داده باید اصلاح شود.
- ناکارآمدی در محیط‌های پویا: در محیط‌هایی با کاربران متغیر عملکرد محدود دارد.

:Open-Set Authentication

- پیچیدگی در پیاده‌سازی: نیاز به الگوریتم‌های پیشرفته برای مدیریت ناشناسان.
- افزایش نرخ خطای مثبت کاذب (False Positive): احتمال اشتباه در طبقه‌بندی کاربران ناشناخته.
- نیاز به منابع محاسباتی بالا: تحلیل داده‌ها و تشخیص ناشناسان ممکن است به قدرت پردازشی بیشتری نیاز داشته باشد.

تکنیک‌های مورد استفاده هر کدام با توجه به کاربرد آن ها:

از تکنیک‌های مورد استفاده در روش Closed-Set Authentication می‌توان به الگوریتم‌های تطبیق الگو (Pattern Matching) و یادگیری با داده‌های کاملاً برچسب‌خورده اشاره کرد. اما در روش Open-Set Authentication بیشتر از تکنیک‌های یادگیری ماشین با قابلیت تشخیص ناشناسان (Outlier Detection)، شبکه‌های عصبی با رویکرد Open-Set Recognition و الگوریتم‌های تحلیل رفتار و تشخیص ناهنجاری (Anomaly Detection) استفاده می‌شود.

چالش‌های استفاده از روش Closed-Set Authentication می‌توان به آسیب‌پذیری در برابر کاربران ناشناخته و تهدیدات جدید و ناکارآمدی در مقیاس‌های بزرگ یا پویا اشاره کرد ولی چالش‌های Open-Set Authentication شامل دشواری در تنظیم مرز دقیق بین کاربران مجاز و غیرمجاز و نیاز به آموزش مدل‌ها با داده‌های متنوع و ناشناخته می‌شود.

پس به طور کلی روش Closed-Set Authentication مناسب سیستم‌هایی است که کاربران آن از پیش تعیین‌شده و ثابت هستند. این روش در محیط‌های کنترل‌شده و با داده‌های مشخص عملکرد بهتری دارد. در مقابل، Open-Set Authentication گزینه‌ای ایده‌آل برای محیط‌های پویا، مقیاس‌پذیر و با کاربران متغیر است، زیرا قابلیت شناسایی و مدیریت ناشناسان را دارد.

بررسی چگونگی پیاده سازی این دو روش

مراحل پیاده سازی Closed-set authentication:

ابتدا مراحل کلی آورده شده سپس در آخر به بررسی الگوریتم‌های پیاده سازی برای این روش می‌پردازیم.

1. جمع‌آوری داده‌های اولیه

برای شناسایی دقیق کاربران، داده‌های بیومتریک (اثر انگشت، چهره، صدا) یا اطلاعات رفتاری (مانند الگوی تایپ) جمع‌آوری می‌شود. این مرحله پایه‌ای برای ایجاد سیستم شناسایی است.

2. ساخت پایگاه داده کاربران

داده‌های جمع‌آوری‌شده در پایگاه داده‌ای ذخیره می‌شود که باید رمزنگاری‌شده و از دسترسی‌های غیرمجاز محافظت گردد.

3. انتخاب و پیاده‌سازی الگوریتم‌های تطبیق

انتخاب الگوریتم تطبیق بستگی به نوع داده‌های جمع‌آوری‌شده دارد. به‌عنوان مثال (در ادامه بیشتر در مورد الگوریتم های مورد استفاده توسط این روش توضیح می‌دهیم):

الگوریتم‌های تطبیق اثر انگشت: Minutiae-based Matching

پردازش صدا: DTW(Dynamic Time Warping)

4. توسعه نرم‌افزار احراز هویت

نرم‌افزار مرکزی برای مدیریت داده‌های ورودی و انجام فرآیند شناسایی و تایید طراحی می‌شود. این نرم‌افزار باید به گونه‌ای باشد که داده‌های کاربر را با الگوریتم‌های تطبیق مقایسه کند و نتیجه نهایی را تولید نماید.

5. ایجاد لایه‌های امنیتی

برای جلوگیری از دسترسی‌های غیرمجاز و محافظت از داده‌های حساس، باید از رمزنگاری قوی (AES، RSA) و پروتکل‌های ایمن شبکه استفاده کرد.

6. آزمایش و ارزیابی سیستم

سیستم باید تحت آزمایش‌های جامع برای ارزیابی نرخ شناسایی صحیح (True Positive Rate)، نرخ خطا (False Positive Rate) و عملکرد در شرایط واقعی قرار گیرد.

7. استقرار و پشتیبانی

پس از آزمایش موفقیت‌آمیز، سیستم در محیط عملیاتی مستقر شده و تیمی برای نظارت و پشتیبانی از آن تعیین می‌شود.

برخی الگوریتم‌های پیاده سازی این روش عبارتند از:

1. Pattern-Based Algorithms

1.1. Minutiae-Based Matching

این الگوریتم نقاط خاصی در الگوهای اثر انگشت (مانند شاخه‌ها یا تقاطع‌ها) را شناسایی کرده و این نقاط را با نمونه ذخیره‌شده مقایسه می‌کند. از مزیت‌های آن به سرعت بالا در تطبیق و مناسب بودن برای پایگاه داده‌های بزرگ نام برد.

1.2. DTW (Dynamic Time Warping)

این الگوریتم اختلافات زمانی در بین سیگنال‌های صوتی را جبران می‌کند و تطابق را بر اساس الگوهای صوتی انجام می‌دهد. مزیت آن در دقت بالا در شناسایی گفتار و مناسب برای داده‌های صوتی متغیر است.

2. Machine Learning Algorithms

2.1. SVMs (Support Vector Machines)

این الگوریتم یک مرز تصمیم‌گیری میان کلاس‌ها ایجاد کرده و داده‌های ورودی را با داده‌های شناخته‌شده مقایسه می‌کند. کاربرد آن بیشتر در شناسایی چهره و ویژگی‌های رفتاری است. از مزایای آن می‌توان به کارایی بالا در مسائل چندبعدی و پشتیبانی از داده‌های پیچیده تر اشاره کرد.

2.2. Random Forest

این الگوریتم با ایجاد چندین درخت تصمیم‌گیری و ترکیب نتایج آن‌ها، دقت در شناسایی را افزایش می‌دهد و کاربرد آن در شناسایی ترکیب ویژگی های چندگانه است. مزایای آن شامل مقاومت در برابر داده‌های نویزی و توانایی بالا در مدیریت داده‌های ترکیبی می‌شود.

3. Distance-Based Algorithms

3.1. Euclidean Distance

فاصله بین ویژگی‌های داده ورودی و داده‌های ذخیره‌شده محاسبه می‌شود. اگر فاصله کمتر از یک آستانه خاص باشد، داده پذیرفته می‌شود. این الگوریتم برای مجموعه داده‌های کوچک مناسب است و پیاده سازی ساده ای دارد.

3.2. Cosine Similarity

روش کار این الگوریتم زاویه بین بردارهای ویژگی داده‌های ورودی و ذخیره‌شده محاسبه می‌شود. زاویه کوچک‌تر نشان‌دهنده شباهت بیشتر است.در شناسایی متن و صوت کاربرد دارد و مناسب برای داده‌های چندبعدی است و در برابر مقیاس بندی مقاوم است.

4. Neural Network Algorithms

4.1. Convolutional Neural Networks) CNNs)

CNN از لایه‌های کانولوشن برای استخراج ویژگی‌های مهم تصاویر و مقایسه آن‌ها با نمونه‌های ذخیره‌شده استفاده می‌کند. مزیت آن در دقت بسیار بالا در پردازش تصاویر و قابلیت یادگیری خودکار ویژگی‌ها است.

4.2. Recurrent Neural Networks) RNNs)

RNN از ساختار بازگشتی برای یادگیری داده‌های متوالی و تطبیق آن‌ها برای تشخیص گفتار و صوت استفاده می‌کند. این روش مناسب برای داده‌های ترتیبی و مدیریت اطلاعات زمانی در داده است.

5. Hybrid Algorithms

5.1. Ensemble Methods

در این روش نتایج چندین الگوریتم (مانند SVM و Random Forest) ترکیب می‌شود تا تصمیم نهایی گرفته شود. این باعث کاهش نرخ خطا و انعطاف‌پذیری در محیط‌های پیچیده می‌شود.

مراحل پیاده سازی Open-set authentication:

Open-Set Authentication سیستمی است که علاوه بر شناسایی کاربران ثبت‌شده، توانایی تشخیص کاربران ناشناخته را نیز دارد. این تفاوت کلیدی مستلزم استفاده از تکنیک‌های پیشرفته‌تری در طراحی و پیاده‌سازی سیستم است. در ادامه به شرح مراحل و توضیح تفاوت هر بخش با closed-set authentication می‌پردازیم. همچنین در آخر به برخی از الگوریتم‌های پیاده سازی این روش می‌پردازیم.

1. جمع‌آوری داده‌های اولیه

همانند Closed-Set، ابتدا داده‌های بیومتریک (اثر انگشت، چهره، صدا) یا اطلاعات رفتاری (الگوی تایپ) برای کاربران ثبت‌شده جمع‌آوری می‌شود.

ویژگی خاص Open-Set، داده‌هایی که نمایانگر رفتارهای ناهنجار یا کاربران ناشناخته باشد نیز ممکن است جمع‌آوری شود تا سیستم توانایی شناسایی الگوهای خارج از کلاس را پیدا کند.

2. ساخت پایگاه داده کاربران و مدل‌های نماینده

در Closed-Set، پایگاه داده فقط شامل کاربران ثبت‌شده است.ولی در Open-Set: علاوه بر داده‌های کاربران ثبت‌شده، مدل‌های نمایشی یا متریک‌های فاصله‌ای (مانند One-vs-Rest Classifiers یا Threshold Models) برای شناسایی ناشناسان اضافه می‌شود.

3. استفاده از الگوریتم‌های تشخیص ناشناسان

در Open-Set Authentication، هدف اصلی تشخیص کاربران ناشناخته است. برای این کار، الگوریتم‌های خاصی استفاده می‌شود (در ادامه بیشتر به این الگوریتم ها می‌پردازیم):

- Outlier Detection: شناسایی داده‌هایی که خارج از محدوده رفتارهای عادی کاربران ثبت‌شده قرار دارند.

- Open-Set Classifiers: مانند SVM (Support Vector Machines) با مرزهای مشخص یا Deep Neural Networks با مکانیزم Uncertainty.

4. توسعه نرم‌افزار با تمرکز بر مرزهای ناشناسان

نرم‌افزار باید به‌گونه‌ای طراحی شود که بتواند تصمیم‌گیری در مورد ناشناسان را به‌صورت دینامیک مدیریت کند.

ویژگی خاص Open-Set اضافه کردن Thresholds یا Confidence Levels برای تعیین قطعیت در مورد ناشناسان.

5. ایجاد لایه‌های امنیتی پیشرفته

در Open-Set Authentication، حفاظت از داده‌های حساس اهمیت بیشتری دارد زیرا تشخیص ناشناسان ممکن است با خطا همراه باشد.

ویژگی خاص Open-Set: استفاده از لایه‌های امنیتی چندگانه برای جلوگیری از False Positives در مورد ناشناسان.

6. آزمایش در شرایط واقعی

در Open-Set، آزمایش باید شامل داده‌های ناشناخته نیز باشد:

- TPR (True Positive Rate): درصد شناسایی کاربران ثبت‌شده.
- FAR (False Acceptance Rate): درصد اشتباه در پذیرش ناشناسان.

7. استقرار و پشتیبانی

برای استقرار Open-Set، سیستم باید دائماً به‌روزرسانی شود تا داده‌های جدید ناشناسان را مدیریت کند.

ویژگی خاص Open-Set، اضافه کردن مکانیزم یادگیری مداوم (Continual Learning) است.

برخی الگوریتم‌های پیاده سازی این روش عبارتند از:

1. Open-Set Classifiers

1.1. Support Vector Machines) SVMs
SVM با استفاده از بردارهای پشتیبان مرزی بین داده‌های شناخته‌شده و ناشناخته تعیین می‌کند.
در Open-Set SVM، مرزهای خاصی برای تفکیک داده‌های ناشناخته از شناخته‌شده طراحی می‌شود.
در روش open-set از تابع هزینه غیر خطی برای کاهش مثبت کاذب (False Positive) استفاده می‌شود و همچنین مقادیر Threshold برای دسته‌بندی ناشناسان تنظیم می‌شود.

1.2. OpenMax Classifier
OpenMax، جایگزین لایه نهایی طبقه‌بندهای شبکه عصبی (SoftMax) می‌شود تا ناشناسان را مدیریت کند. این الگوریتم توزیع خروجی‌ها را تحلیل کرده و کلاس "ناشناخته" را در میان کلاس‌های ممکن اضافه می‌کند. مزیت آن در Open-set در امکان اضافه کردن ناشناسان به‌صورت پویا و عملکرد بهینه برای داده‌های با توزیع غیرمعمول است.

2. Outlier Detection Algorithms

2.1. (k-Nearest Neighbors) k-NN
فاصله داده ورودی با نزدیک‌ترین نقاط در فضای ویژگی محاسبه می‌شود. اگر این فاصله از یک مقدار آستانه بیشتر باشد، داده به‌عنوان ناشناس طبقه‌بندی می‌شود. مزیت آن در سادگی در پیاده‌سازی و کارایی برای داده‌های با ابعاد کم است.

2.2. EVT (Extreme Value Theory)

EVT بر اساس تئوری مقادیر حدی، داده‌هایی که فراتر از توزیع نرمال هستند را به‌عنوان ناشناس طبقه‌بندی می‌کند. این روش مناسب برای داده‌های با ابعاد بالا است و مثبت کاذب را با استفاده از تحلیل توزیع کاهش می‌دهد.

3. Deep Learning Approaches

3.1. Autoencoders

Autoencoder مدل‌های مولدی هستند که داده‌های شناخته‌شده را با دقت بازسازی می‌کنند. اگر داده ناشناس باشد، خطای بازسازی افزایش می‌یابد و آن را به‌عنوان ناشناس شناسایی می‌کند. از مزیت های این روش مناسب بودن برای داده‌های پیچیده وچندبعدی و قابلیت یادگیری ویژگی‌های خاص داده‌ها است.

3.2. Generative Adversarial Networks (GANs)

GAN از دو شبکه (مولد و تمایز دهنده) استفاده می‌کند. تمایز دهنده داده‌های ناشناخته را که توسط مولد تولید نشده‌اند، شناسایی می‌کند. مزیت های این روش در قدرت بالا در تشخیص الگوهای غیر معمول و مناسب بودن برای مجموعه داده‌های متنوع است.

4. Hybrid Models

4.1. One-vs-Rest Classifiers

یک طبقه‌بند برای هر کلاس شناخته‌شده ایجاد می‌شود. داده‌ای که با هیچ‌کدام از این طبقه‌بندها همخوانی ندارد، ناشناس شناخته می‌شود. مزیت آن در ساده و قابل پیاده‌سازی برای مجموعه داده‌های کوچک و عملکرد بالا در تنظیمات خاص است.

4.2. Threshold-Based Models

مقادیر خروجی مدل‌ها (مانند نمره اطمینان) با آستانه مقایسه می‌شود. اگر مقدار کمتر از آستانه باشد،

داده ناشناس در نظر گرفته می‌شود. مزیت آن در سادگی در پیاده‌سازی. مناسب بودن برای تنظیمات آنلاین و بلادرنگ است.

در مورد روش SVM چون در هر دو روش closed-set و open-set آمده است و بعدا در بخش پیاده سازی قرار است از آن استفاده کنیم کمی بیشتر توضیح می‌دهم.

ماشین بردار پشتیبان (**Support Vector Machine** یا **SVM**) الگوریتمی در حوزه یادگیری ماشین نظارت‌شده است که برای مسائل طبقه‌بندی و رگرسیون به کار می‌رود. هدف اصلی SVM یافتن یک **ابرصفحه** (Hyperplane) است که داده‌های متعلق به دسته‌های مختلف را با حداکثر حاشیه از هم جدا کند.

نحوه کارکرد SVM:

1. جداسازی خطی: در مواردی که داده‌ها به صورت خطی قابل جداسازی هستند، SVM ابرصفحه‌ای را می‌یابد که بیشترین فاصله (حاشیه) را بین نزدیک‌ترین نمونه‌های هر دسته ایجاد کند. این نمونه‌های مرزی به عنوان **بردارهای پشتیبان** شناخته می‌شوند و تعیین‌کننده موقعیت ابرصفحه هستند.

2. جداسازی غیرخطی: در بسیاری از مسائل واقعی، داده‌ها به صورت خطی قابل جداسازی نیستند. در این حالت، SVM با استفاده از **توابع کرنل** (Kernel Functions) داده‌ها را به فضایی با ابعاد بالاتر نگاشت می‌کند تا در آن فضا جداسازی خطی ممکن شود. توابع کرنل متداول عبارت‌اند از:

کرنل چندجمله‌ای (Polynomial Kernel): برای مدل‌سازی روابط غیرخطی با درجات مختلف.

کرنل تابع پایه شعاعی (RBF): برای جداسازی داده‌هایی که مرزهای پیچیده و غیرخطی دارند.

کرنل سیگموئید (Sigmoid Kernel): مشابه عملکرد توابع فعال‌سازی در شبکه‌های عصبی.

حاشیه نرم (Soft Margin): در مواردی که داده‌ها به طور کامل قابل جداسازی نیستند یا دارای نویز هستند، SVM از مفهوم حاشیه نرم استفاده می‌کند که اجازه می‌دهد برخی نمونه‌ها در سمت نادرست

ابرصفحه قرار گیرند، اما با افزودن یک جریمه به تابع هدف، تلاش می‌کند تعداد این نمونه‌ها را به حداقل برساند.

مزایای SVM:

- **کارایی در فضاهای با ابعاد بالا:** SVM در داده‌هایی با ابعاد بالا عملکرد خوبی دارد و می‌تواند با استفاده از توابع کرنل، جداسازی‌های پیچیده را انجام دهد.
- **مقاومت در برابر بیش‌برازش (Overfitting):** با انتخاب مناسب پارامترهای مدل و تابع کرنل، SVM می‌تواند از بیش‌برازش جلوگیری کند، به‌ویژه در مسائلی که تعداد ویژگی‌ها بیشتر از تعداد نمونه‌هاست.

معایب SVM:

- **پیچیدگی محاسباتی:** در مجموعه داده‌های بسیار بزرگ، آموزش SVM می‌تواند زمان‌بر باشد و به منابع محاسباتی بالایی نیاز داشته باشد.
- **انتخاب تابع کرنل مناسب:** انتخاب نادرست تابع کرنل می‌تواند به کاهش دقت مدل منجر شود؛ بنابراین، نیاز به تجربه و دانش در انتخاب کرنل مناسب است.

بررسی کاربرد های آن در voice authentication

Closed-Set Voice Authentication:

همانطور که پیش‌تر اشاره شد، در این رویکرد، سیستم تنها قادر به شناسایی و تایید هویت کاربرانی است که قبلاً در پایگاه داده ثبت شده‌اند. به عبارت دیگر، اگر صدای ورودی متعلق به یکی از کاربران شناخته‌شده نباشد، سیستم نمی‌تواند هویت او را تشخیص دهد.

کاربردها:

- کنترل دسترسی به سیستم‌های حساس: در سازمان‌ها و نهادهایی که نیاز به امنیت بالایی دارند، از Closed-Set Voice Authentication برای اطمینان از دسترسی فقط افراد مجاز به اطلاعات حساس استفاده می‌شود.
- شناسایی گوینده در مراکز تماس: مراکز تماس می‌توانند با استفاده از این فناوری، هویت مشتریان را از طریق صدای آن‌ها تأیید کرده و خدمات شخصی‌سازی‌شده ارائه دهند.
- دستگاه‌های هوشمند خانگی: دستگاه‌هایی مانند اسپیکرهای هوشمند می‌توانند با استفاده از Closed-Set Voice Authentication، دستورات را تنها از کاربران مجاز پذیرفته و اجرا کنند، که این امر به افزایش امنیت و جلوگیری از دسترسی غیرمجاز کمک می‌کند.

Open-Set Voice Authentication:

در این رویکرد، سیستم نه تنها قادر به شناسایی کاربران ثبت‌شده است، بلکه می‌تواند تشخیص دهد که صدای ورودی متعلق به فردی خارج از مجموعه کاربران شناخته‌شده است. این ویژگی امکان شناسایی کاربران جدید یا تشخیص نفوذهای احتمالی را فراهم می‌کند.

کاربردها:

- سیستم‌های امنیتی پیشرفته: در مکان‌هایی که نیاز به تشخیص نفوذگران یا افراد غیرمجاز است، Open-Set Voice Authentication می‌تواند با شناسایی صداهای ناشناخته، امنیت را افزایش دهد.
- مراکز تماس با حجم بالای مشتریان: در مراکزی که امکان ثبت صدای تمامی مشتریان وجود ندارد، این سیستم می‌تواند با تشخیص صداهای جدید، فرآیند احراز هویت را تسهیل کند.
- کاربردهای قضایی و قانونی: در تحلیل‌های صوتی مرتبط با پرونده‌های قضایی، Open-Set Voice Authentication می‌تواند به شناسایی افراد ناشناخته در ضبط‌های صوتی کمک کند.

تفاوت‌ها:

- محدوده شناسایی: در Closed-Set، سیستم تنها کاربران ثبت‌شده را شناسایی می‌کند، در حالی که در Open-Set، سیستم قادر به تشخیص صداهای ناشناخته نیز هست.
- پیچیدگی پیاده‌سازی: Open-Set Voice Authentication به دلیل نیاز به تشخیص صداهای ناشناخته، پیچیدگی بیشتری در پیاده‌سازی دارد.
- کاربردها: Closed-Set بیشتر در محیط‌های با کاربران محدود و شناخته‌شده کاربرد دارد، در حالی که Open-Set در محیط‌هایی با تعداد کاربران نامشخص یا متغیر مفید است.

چالش‌های voice authentication

شناسایی و توضیح چالش های اصلی که در تحقیقات و کاربردهای authentication voice و gender classification وجود دارد و بررسی راه حل های بالقوه و تحقیقات جاری برای غلبه بر این چالش ها

در حوزه‌های احراز هویت صوتی (Voice Authentication) و تشخیص جنسیت از روی صدا (Gender Classification)، چالش‌های متعددی وجود دارد که می‌تواند بر دقت و کارایی سیستم‌ها تأثیر بگذارد. در ادامه، به برخی از این چالش‌ها و راه حل های آن ها پرداخته می‌شود:

چالش‌های احراز هویت صوتی

تغییرات در صدای کاربر: صدای هر فرد می‌تواند به دلایلی مانند بیماری، خستگی، استرس یا افزایش سن تغییر کند. این تغییرات می‌توانند بر الگوهای صوتی تأثیر گذاشته و دقت سیستم‌های احراز هویت صوتی را کاهش دهند. برخی راه حل هایی که برای این چالش وجود دارد عبارتند از:

استفاده از الگوریتم‌های یادگیری عمیق که قادر به مدل‌سازی تغییرات طبیعی در صدای افراد هستند، می‌تواند به بهبود دقت سیستم‌های احراز هویت صوتی کمک کند.

توسعه مدل‌های مقاوم در برابر تغییرات صدا که با استفاده از داده‌های متنوع آموزشی، توانایی شناسایی کاربران را حتی در شرایطی که صدای آن‌ها تغییر کرده است، داشته باشند.

نویز و شرایط محیطی: نویزهای محیطی مانند صداهای مزاحم می‌توانند نقاط اشتباه و خطا در تشخیص هویت ایجاد کرده و دقت سیستم را کاهش دهند. برخی راه حل هایی که برای این چالش وجود دارد عبارتند از:

بهره‌گیری از فیلترهای حذف نویز پیشرفته و تکنیک‌های پیش‌پردازش سیگنال می‌تواند به کاهش تأثیر نویزهای محیطی بر دقت سیستم‌های احراز هویت صوتی کمک کند.

استفاده از مدل‌های یادگیری ماشین که برای کار در محیط‌های نویزی آموزش دیده‌اند، می‌تواند مقاومت سیستم را در برابر نویز افزایش دهد. [14]

امنیت در برابر جعل صدا (Spoofing): تکنیک‌هایی مانند استفاده از صدای ضبط‌شده یا تولید صدای تقلبی با کمک فناوری‌های پیشرفته می‌توانند سیستم‌های احراز هویت صوتی را فریب دهند. این مسئله امنیت سیستم را به چالش می‌کشد. برخی راه حل هایی که برای این چالش وجود دارد عبارتند از:

توسعه الگوریتم‌های تشخیص تقلب که قادر به شناسایی صداهای مصنوعی یا تقلیدی هستند، می‌تواند امنیت سیستم‌های احراز هویت صوتی را افزایش دهد. [15]

استفاده از ویژگی‌های بیومتریک چندگانه مانند ترکیب تشخیص صدا با تشخیص چهره یا اثر انگشت، می‌تواند مقاومت سیستم را در برابر حملات جعل صدا افزایش دهد.

چالش‌های مقیاس‌پذیری: در سیستم‌هایی با تعداد بالای کاربران، سرعت و دقت شناسایی ممکن است کاهش یابد. این مسئله نیازمند بهینه‌سازی الگوریتم‌ها و زیرساخت‌های مناسب است. برخی راه حل هایی که برای این چالش وجود دارد عبارتند از:

توسعه الگوریتم‌های بهینه و کارآمد که قادر به پردازش داده‌های بزرگ با سرعت و دقت بالا هستند، می‌تواند به حل مشکلات مقیاس‌پذیری کمک کند.

استفاده از زیرساخت‌های محاسباتی پیشرفته مانند رایانش ابری می‌تواند به مدیریت حجم بالای داده‌ها و کاربران کمک کند.

چالش‌های تشخیص جنسیت از روی صدا

ویژگی‌های غیرمشخص صدا: برخی صداها به‌طور طبیعی دارای ویژگی‌های مبهم یا مشترک میان جنسیت‌ها هستند که باعث کاهش دقت سیستم می‌شود. این مسئله به‌ویژه در افرادی با صداهای میان‌جنسیتی یا تغییر جنسیت داده‌شده مشهود است. برخی راه حل هایی که برای این چالش وجود دارد عبارتند از:

استفاده از الگوریتم‌های یادگیری عمیق که قادر به استخراج ویژگی‌های پیچیده و نامحسوس صدا هستند، می‌تواند به بهبود دقت تشخیص جنسیت کمک کند. [13]

توسعه مدل‌های ترکیبی که از اطلاعات چندگانه مانند صدا و تصویر بهره می‌برند، می‌تواند دقت تشخیص جنسیت را افزایش دهد.

تأثیر زبان و گویش: تفاوت‌های زبانی، گویشی یا لهجه‌ها می‌توانند بر الگوهای صوتی تأثیر بگذارند و شناسایی جنسیت را دشوارتر کنند. این مسئله نیازمند مدل‌هایی است که قادر به تطبیق با گویش‌ها و زبان‌های مختلف باشند. برخی راه حل هایی که برای این چالش وجود دارد عبارتند از:

جمع‌آوری داده‌های آموزشی متنوع از زبان‌ها و گویش‌های مختلف می‌تواند به مدل‌ها کمک کند تا با تنوع زبانی سازگار شوند.

استفاده از مدل‌های چندزبانه که قادر به پردازش و تحلیل صداهای مربوط به زبان‌ها و گویش‌های مختلف هستند، می‌تواند به بهبود دقت تشخیص جنسیت کمک کند.

عدم توازن داده‌ها: در بسیاری از موارد، داده‌های آموزش برای جنسیت‌های مختلف (مرد/زن) به‌طور مساوی توزیع نشده‌اند. این عدم توازن می‌تواند به کاهش دقت مدل‌ها منجر شود. برخی راه حل هایی که برای این چالش وجود دارد عبارتند از:

استفاده از تکنیک‌های افزایش داده (Data Augmentation) برای ایجاد تعادل در مجموعه داده‌ها می‌تواند به بهبود عملکرد مدل‌ها کمک کند.

جمع‌آوری داده‌های بیشتر از گروه‌های کم‌نماینده می‌تواند به ایجاد توازن در داده‌ها و بهبود دقت مدل‌ها کمک کند.

چالش‌های مشترک

افزایش دقت در شرایط دنیای واقعی: سیستم‌ها در محیط‌های کنترل‌شده عملکرد خوبی دارند، اما در شرایط واقعی با نویز، گویش‌های متفاوت و تغییرات صدا چالش‌های جدی دارند. برخی راه حل هایی که برای این چالش وجود دارد عبارتند از:

استفاده از الگوریتم‌های یادگیری عمیق که قادر به مدل‌سازی پیچیدگی‌های موجود در داده‌های صوتی واقعی هستند، می‌تواند به بهبود دقت سیستم‌ها کمک کند. [16]

توسعه مدل‌های مقاوم در برابر نویز که با استفاده از داده‌های متنوع آموزشی، توانایی شناسایی و تشخیص را حتی در حضور نویز های محیطی داشته باشند.

کمبود داده‌های متنوع: بسیاری از مجموعه داده‌های موجود تنوع کافی ندارند، به‌ویژه برای گویش‌های مختلف، صداهای میان‌جنسیتی یا افراد با تغییرات صدای موقتی. برخی راه حل هایی که برای این چالش وجود دارد عبارتند از:

جمع‌آوری داده‌های متنوع و جامع از کاربران مختلف با شرایط گوناگون می‌تواند به بهبود عملکرد مدل‌ها کمک کند.

استفاده از تکنیک‌های افزایش داده (Data Augmentation) برای ایجاد تعادل در مجموعه داده‌ها می‌تواند به بهبود عملکرد مدل‌ها کمک کند.

حریم خصوصی و امنیت داده‌ها: ذخیره‌سازی و پردازش داده‌های صوتی کاربران ممکن است نگرانی‌هایی درباره حریم خصوصی ایجاد کند. این مسئله نیازمند رعایت استانداردهای امنیتی و اخلاقی است. برخی راه حل هایی که برای این چالش وجود دارد عبارتند از:

استفاده از روش‌های رمزنگاری پیشرفته برای حفاظت از داده‌های صوتی کاربران می‌تواند به حفظ حریم خصوصی کمک کند.

توسعه سیاست‌های دسترسی محدود به داده‌های حساس می‌تواند از سوءاستفاده‌های احتمالی جلوگیری کند.

هزینه پردازش و زیرساخت‌ها: پیاده‌سازی سیستم‌های دقیق با استفاده از الگوریتم‌های پیچیده می‌تواند منابع محاسباتی زیادی مصرف کند. این مسئله به‌ویژه در کاربردهای بلادرنگ اهمیت دارد. برخی راه حل هایی که برای این چالش وجود دارد عبارتند از:

استفاده از زیرساخت‌های محاسباتی پیشرفته مانند رایانش ابری می‌تواند به مدیریت حجم بالای داده‌ها و کاربران کمک کند.

توسعه الگوریتم‌های بهینه و کارآمد که قادر به پردازش داده‌های بزرگ با سرعت و دقت بالا هستند، می‌تواند به حل مشکلات مقیاس‌پذیری کمک کند. [17]

اهمیت پیش پردازش داده های صوتی در زمینه voice authentication و gender classification:

1. **دقت بهبود یافته**

پیش‌پردازش داده‌های صوتی دقت سیستم‌های تشخیص هویت صوتی و طبقه‌بندی جنسیت را با اطمینان از تمیزی و مرتبط بودن سیگنال‌های ورودی افزایش می‌دهد [6]. نویز محیطی، مانند صداهای پس‌زمینه یا اعوجاج‌ها، می‌تواند توانایی مدل در استخراج الگوهای معنادار از داده‌ها را مختل کند. حذف این نویز باعث می‌شود مدل فقط صدای گوینده را پردازش کند.

به‌عنوان مثال:

- در طبقه‌بندی جنسیت، تغییرات جزئی در فرکانس زیر و فرکانس‌های فورمنت نشانه‌های حیاتی برای تعیین جنسیت هستند. نویز می‌تواند این ویژگی‌ها را مخفی کند و منجر به خطای طبقه‌بندی شود.
- در احراز هویت صوتی، تطابق دقیق ویژگی‌های صوتی با پروفایل ذخیره‌شده ضروری است. بدون پیش‌پردازش، خطاهای ناشی از اعوجاج‌های موجود در داده‌ها می‌توانند به احراز هویت یا رد اشتباه منجر شوند.

2. **بهبود ویژگی‌ها**

داده‌های خام صوتی اطلاعات زیادی دارند که بسیاری از آن‌ها برای وظایف مورد نظر بی‌ربط هستند. پیش‌پردازش بر استخراج و تقویت ویژگی‌هایی تمرکز دارد که بیشترین اطلاعات مفید را برای کاربرد هدف فراهم می‌کنند.

ویژگی‌های کلیدی مانند:

- MFCCs که نمایانگر طیف توان صدا بوده و ویژگی‌های صوتی و تمبر صوتی را به‌خوبی ثبت می‌کنند.
- فرکانس زیر و فورمنت‌ها که ویژگی‌های خاص جنسیت و گوینده را نشان می‌دهند.

- سطوح انرژی که می‌توانند بخش‌های گفتار و سکوت را شناسایی کنند و به بخش‌بندی کمک کنند.

این ویژگی‌ها به مدل‌های یادگیری ماشین کمک می‌کنند تا ویژگی‌های منحصربه‌فرد صداها را برای احراز هویت و طبقه‌بندی تشخیص دهند.

3. **کارایی محاسباتی**

داده‌های صوتی پردازش‌نشده بسیار پیچیده و دارای ابعاد بالا هستند. آموزش مدل‌ها با این داده‌ها نیازمند منابع پردازشی و حافظه بیشتری است. پیش‌پردازش داده‌ها را ساده می‌کند از طریق:

- کاهش افزونگی و فشرده‌سازی اطلاعات به نمایش‌های کوچکتر و قابل مدیریت‌تر.
- استانداردسازی فرمت ورودی که پردازش آن را برای مدل‌ها آسان‌تر می‌کند.

پیش‌پردازش مؤثر منجر به آموزش سریع‌تر مدل‌ها و پیش‌بینی‌های بلادرنگ می‌شود که برای کاربردهایی مانند احراز هویت صوتی موبایلی و طبقه‌بندی جنسیت ضروری است[9].

4. **مقاومت در برابر تغییرات**

گفتار انسان به دلایلی مانند لهجه ها، سرعت صحبت کردن و شرایط ضبط ممکن است کاملا یکتا نباشد. پیش پردازش هایی مانند نرمال‌سازی و مقیاس‌بندی فرکانس زیر این تغییرات را کاهش می‌دهند. با نرمال‌سازی دامنه، متعادل‌سازی دامنه فرکانسی، و حذف بخش‌های سکوت یا بی‌ربط، پیش‌پردازش تضمین می‌کند که مدل‌ها بر اساس داده‌های سازگار آموزش ببینند. این امر سیستم را در برابر شرایط واقعی که تغییرات اجتناب‌ناپذیر است، مقاوم‌تر می‌کند.

5. **امنیت پیشرفته**

در سیستم‌های احراز هویت صوتی، پیش‌پردازش علاوه بر افزایش دقت، امنیت را نیز تقویت می‌کند. تلاش‌های تقلبی، مانند سنتز صدا یا حملات بازپخش، اغلب دارای بی‌نظمی‌هایی هستند که پیش‌پردازش می‌تواند آن‌ها را تشخیص دهد. تکنیک‌هایی مانند تحلیل طیفی می‌توانند الگوهای غیرطبیعی را نشان دهند، در حالی که کاهش نویز تضمین می‌کند که ویژگی‌های واقعی صدا حفظ شوند.

پیش‌پردازش مؤثر، سیستم‌های احراز هویت صوتی را در برابر دستکاری مقاوم‌تر کرده و امنیت داده‌های حساس را تضمین می‌کند و قابلیت اطمینان در برنامه‌های کاربردی با امنیت بالا را افزایش می‌دهد.

توضیح مراحل پیش پردازش

همانطور که اشاره شد، پیش‌پردازش گامی حیاتی در آماده‌سازی داده‌های صوتی برای کاربردهایی مانند احراز هویت صوتی و تشخیص جنسیت است. این فرآیند شامل تبدیل سیگنال‌های صوتی خام به یک قالب تمیز، استاندارد و غنی از ویژگی‌هاست که عملکرد مدل را بهبود می‌بخشد. از روش‌های معمول پیش‌پردازش می‌توان به کاهش نویز، نرمال‌سازی، پنجره‌بندی، استخراج ویژگی‌ها (مانند MFCC، ویژگی‌های طیفی)، بازنمونه‌گیری و حذف سکوت اشاره کرد. در این بخش به سه تکنیک اساسی پرداخته شده است: کاهش نویز، نرمال‌سازی و پنجره‌بندی.

۱. کاهش نویز

کاهش نویز به فرآیند حذف صداهای پس‌زمینه یا تداخل‌های ناخواسته از سیگنال صوتی گفته می‌شود. نویز محیطی مانند صدای گفتگو، باد یا نویز الکترونیکی می‌تواند ویژگی‌های حیاتی سیگنال صوتی را مبهم کند و به نادرستی در تحلیل منجر شود.

از اهمیت های آن میتوان به افزایش وضوح سیگنال با بالا بردن نسبت سیگنال به نویز (SNR)، بهبود استخراج ویژگی با حفظ ویژگی های زیر و بمی و پایداری سیستم در محیط های واقعی با سطوح نویز متغیر اشاره کرد.

تکنیک‌ها

- **گیتینگ طیفی:**

فرکانس‌های ضعیف نویز را حذف می‌کند و ویژگی‌های اصلی گفتار را حفظ می‌کند. فرمول آن:

$$X(f) = \begin{cases} X(f) & \text{if } |X(f)| > T \\ 0 & \text{if } |X(f)| \leq T \end{cases}$$

که در آن X تبدیل فوریه سیگنال و T آستانه نویز است.

- **فیلتر سازی تطبیقی:**

با استفاده از ویژگی‌های لحظه‌ای سیگنال، مانند فیلتر وینر، پارامترهای فیلتر را برای حذف نویز بهینه می‌کند.

- **کاهش نویز مبتنی بر یادگیری عمیق:**

شبکه‌های عصبی مانند U-Net صدای تمیز را از ورودی نویزی پیش‌بینی می‌کنند و کاهش نویز در محیط‌های پیچیده را ارائه می‌دهند. [8]

تأثیر عملی

کاهش نویز باعث می‌شود مدل‌ها بر ویژگی‌های صوتی گوینده تمرکز کنند و از تداخل‌های پس‌زمینه دوری کنند، که منجر به احراز هویت صوتی و تشخیص جنسیت قابل‌اعتمادتر می‌شود.

۲. نرمال‌سازی

نرمال‌سازی دامنه سیگنال‌های صوتی را به سطحی ثابت تنظیم می‌کند و اختلافات ناشی از شرایط یا دستگاه‌های ضبط مختلف را کاهش می‌دهد.

اهمیت آن نیز در زمینه یکپارچگی سیگنال (با اطمینان از قابل مقایسه بودن بلندی همه نمونه های سیگنال)، جلوگیری از پرش سیگنال و بهبود استخراج ویژگی هایی که به انرژی دامنه حساس هستند می‌باشد.

تکنیک‌ها

- **نرمال‌سازی اوج:**

سیگنال را به گونه ای تغییر میدهد که نقطه اوج آن به یک نقطه اوج هدف برسد:

$$x_{\text{normalized}}(t) = x(t) \times \frac{\text{Target Peak}}{\text{Current Peak}}$$

- **نرمال‌سازی RMS:**

سیگنال را به گونه ای تغییر میدهد که به سطح انرژی RMS هدف برسد. ضریب مقیاس آن نیز به صورت زیر است:

$$\text{Scale Factor} = \frac{\text{Target RMS}}{\sqrt{\frac{1}{N}\sum_{n=1}^{N} x(n)^2}}$$

که در آن N تعداد سمپل ها و x(n) مقدار سیگنال در سمپل n ام می‌باشد.

- **فشرده‌سازی دامنه دینامیکی:**

تفاوت بین بلندترین و آرام‌ترین بخش‌های سیگنال را کاهش می‌دهد و یکنواختی را افزایش می‌دهد.

تأثیر عملی

نرمال‌سازی اطمینان حاصل می‌کند که ویژگی‌های ظریف گفتار تحت‌تأثیر تغییرات حجم قرار نمی‌گیرند و مدل‌ها الگوها را مؤثرتر شناسایی می‌کنند.

۳. پنجره‌بندی

پنجره‌بندی سیگنال صوتی پیوسته را به فریم‌های کوچک و همپوشان تقسیم می‌کند. سیگنال‌های گفتاری غیر ایستا هستند، به این معنا که ویژگی‌های آن‌ها در طول زمان تغییر می‌کنند. پنجره‌بندی امکان تحلیل در بازه‌های زمانی کوتاه و تقریباً ایستا را فراهم می‌کند.

این تکنیک اهمیت خاصی در استخراج ویژگی های دینامیکی (منحنی های زیر و بمی و انتقال های طیفی)، کاهش اثرات لبه با نرم کردن انتقال بین فریم ها و تقریب ایستایی دارد.

تکنیک‌ها

- **پنجره مستطیلی:**

این پنجره ساده‌ترین و ابتدایی‌ترین نوع از پنجره‌هاست که معمولاً به عنوان پیش‌فرض در بسیاری از مسائل پردازش سیگنال استفاده می‌شود. در این نوع پنجره، مقادیر وزن‌ها در طول زمان ثابت هستند، به این معنی که همه نقاط سیگنال به طور یکسان و بدون تغییر وزن‌دهی می‌شوند. این سادگی در محاسبات باعث می‌شود که اجرای آن سریع و آسان باشد، اما به دلیل نداشتن خاصیت‌های نرم‌کننده، در تجزیه و تحلیل طیفی موجب نشت طیفی قابل توجهی

می‌شود. نشت طیفی به این معناست که سیگنال‌های غیرضروری به فرکانس‌های دیگر منتقل می‌شوند و موجب کاهش دقت در تحلیل‌های فرکانسی می‌گردد.

- **پنجره هَمینگ:**

این پنجره برای کاهش نشت طیفی طراحی شده و تعادلی بین سرکوب لوب‌های جانبی و حفظ وضوح لوب اصلی برقرار می‌کند. با اعمال ضرب‌کننده کسینوسی، نشت طیفی کمتر می‌شود، ولی لوب اصلی پهن‌تر می‌شود که باعث کاهش دقت در تحلیل‌های فرکانسی می‌شود. فرمول آن:

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right)$$

- **پنجره هَنینگ:**

شبیه پنجره هَمینگ است، اما با تمرکز بیشتر روی کاهش ناپیوستگی‌ها در لبه‌ها. این پنجره نشت طیفی کمتری دارد، اما باز هم دقت فرکانسی تحت تاثیر قرار می‌گیرد، چون لوب اصلی کمی پهن‌تر می‌شود. فرمول آن:

$$w(n) = 0.5\left(1 - \cos\left(\frac{2\pi n}{N-1}\right)\right)$$

- **پنجره بلکمن:**

این پنجره با استفاده از ترکیب چند هارمونیک کسینوسی، نشت طیفی را به شدت کاهش می‌دهد. اما این کار به قیمت کاهش دقت در تحلیل فرکانسی تمام می‌شود، چرا که لوب اصلی گسترش یافته و وضوح فرکانسی کاهش می‌یابد. فرمول آن:

$$w(n) = 0.42 - 0.5 \cos\left(\frac{2\pi n}{N-1}\right) + 0.08 \cos\left(\frac{4\pi n}{N-1}\right)$$

(در این مقاله، هر سه پنجره همینگ، هنینگ و بلکمن مقایسه شده اند[12].)

تأثیر عملی

پنجره‌بندی استخراج ویژگی‌های حساس به زمان مانند MFCC یا فرمنت‌ها را ممکن می‌سازد. برای مثال، در تشخیص جنسیت، فریم‌های کوتاه به ثبت تغییرات زیر و بمی و محتوای طیفی کمک می‌کنند.

تکنیک‌های استخراج ویژگی

استخراج ویژگی یکی از مهم‌ترین مراحل در تجزیه و تحلیل داده‌های صوتی است که تأثیر قابل‌توجهی بر عملکرد الگوریتم‌های یادگیری ماشین دارد. داده‌های صوتی حاوی اطلاعات خام بسیاری هستند، اما تنها بخشی از این اطلاعات برای مسائل خاص (مانند تشخیص گفتار، شناسایی احساسات، یا پردازش موسیقی) مفید هستند. هدف از استخراج ویژگی، کاهش ابعاد داده و تمرکز بر جنبه‌های مفید آن برای مدل‌های یادگیری ماشین است.

1. Fast Fourier Transform

توضیح اولیه و کلی

تبدیل فوریه سریع (FFT) یکی از مهم‌ترین ابزارها در تحلیل داده‌های صوتی است که برای استخراج اطلاعات فرکانسی از سیگنال صوتی استفاده می‌شود. این تکنیک، پایه بسیاری از روش‌های استخراج ویژگی صوتی مانند MFCC و Spectrogram است و به همین دلیل آن را اول توضیح میدهیم. این روش سیگنال را از دامنه زمان به دامنه فرکانس منتقل می‌کند و به ما امکان می‌دهد اجزای فرکانسی سیگنال و شدت آن‌ها را درک کنیم.

توضیح گام به گام و دقیق استخراج این ویژگی

ZERO. پیش‌پردازش: سیگنال صوتی ورودی برای حذف نویز و یکنواخت‌سازی سطح صدا پیش‌پردازش می‌شود. مراحل اختیاری شامل مواردی مثل بازنمونه‌برداری (Resampling) و تنظیم نرخ نمونه‌برداری به یک مقدار ثابت (مثلاً ۴۴.۱ کیلوهرتز) یا مثلا اینکه اگر سیگنال به صورت استریو است، به مونو تبدیل شود، میشوند.

I. سیگنال صوتی یک داده پیوسته است که در طول زمان تغییر می‌کند. برای تحلیل دقیق، سیگنال به فریم‌های کوچک تقسیم می‌شود. طول معمول هر فریم بین ۲۰ تا ۴۰ میلی‌ثانیه است. این تقسیم‌بندی باعث می‌شود بتوان ویژگی‌های سیگنال را در بازه‌های زمانی ایستا

(Stationary) استخراج کرد. تعداد نمونه‌ها در هر فریم از حاصلضرب sample rate برحسب هرتز در طول فریم بدست می‌آید و همپوشانی بین فریم‌ها معمولاً ۵۰٪ تا ۷۵٪ است تا انتقال بین فریم‌ها نرم باشد. نهایتا هم این فریم‌ها با استفاده از یک پنجره (مانند **پنجره هامینگ**) هموار می‌شوند.

:Hamming window

$$w[n] = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right)$$

در اینجا N تعداد نمونه های فریم است و n هم index نمونه است و سیگنال پنجره گذاری شده ی نهایی به فرم $x_w[n] = x[n]\cdot w[n]$ خواهد بود.

II. برای تحلیل فرکانسی، سیگنال پنجره‌گذاری شده به دامنه فرکانس تبدیل می‌شود. این تبدیل با استفاده از تبدیل فوریه سریع (FFT) انجام می‌شود:

$$X[k] = \sum_{n=0}^{N-1} x_w[n] e^{-j\frac{2\pi kn}{N}}$$

در اینجا X[k] دامنه فرکانسی سیگنال است و N تعداد نمونه‌ها در فریم بوده و k همان index فرکانس است.

اما استفاده از FFT چه فوایدی دارد و چرا از آن استفاده می کنیم؟ برای این منظور در همین گام دلایل این کار با بیان چند مقدمه از پایه بیان می شود:

- تعریف تبدیل فوریه گسسته (DFT):

$$X[k] = \sum_{n=0}^{N-1} x[n] \cdot e^{-j\frac{2\pi kn}{N}}, \quad k = 0, 1, \ldots, N-1$$

- در اینجا X[k] دامنه فرکانسی سیگنال است و x[n] دامنه ی زمانی سیگنال است. N تعداد نمونه‌ها در فریم بوده و k همان index فرکانس است. نهایتا $e^{-j2\pi kn/N}$ هسته نمایی که به عنوان ضرایب چرخشی (Twiddle Factors) شناخته می‌شود.
- محاسبه مستقیم DFT نیاز به $O(N^2)$ عملیات دارد، زیرا برای هر یک از k مقادیر X[k]، باید N ضرب و جمع انجام شود.

- **هدف تبدیل فوریه سریع (FFT) : این روش** هدفش کاهش پیچیدگی محاسباتی DFT از $O(N^2)$ به $O(N.log(N))$ است [1]. این بهبود از طریق استفاده از خواص زیر به دست می‌آید ($W^k_N = e^{-j2\pi k/N}$) :

- تقارن نمایی:

$$W_N^{k+N/2} = -W_N^k$$

- دوره‌ای بودن:

$$W_N^{k+N} = W_N^k$$

- هویت بازگشتی: تقسیم یک مسئله DFT با طول N به دو مسئله DFT با طول N/2 میتواند انجام شود.

- **الگوریتم Cooley-Tukey برای FFT :** الگوریتم Cooley-Tukey متداول‌ترین روش برای اجرای FFT است. این الگوریتم از رویکرد تقسیم و غلبه (Divide and Conquer) استفاده

می‌کند. در این روش یک سیگنال x[n] به دو زیر سیگنال تقسیم می‌شود: یکی شامل نمونه‌هایی با ایندکس زوج و یکی شامل نمونه‌های فرد.

حالا DFT کلی با طول N به صورت زیر تعریف می‌شود:

$$X[k] = X_{\text{even}}[k] + W_N^k \cdot X_{\text{odd}}[k]$$

$$X[k + \frac{N}{2}] = X_{\text{even}}[k] - W_N^k \cdot X_{\text{odd}}[k]$$

پس کافیست طی سه مرحله عمل کنیم: ابتدا محاسبه DFT برای نمونه‌های زوج. سپس محاسبه DFT برای نمونه‌های فرد، و نهایتا ترکیب نتایج برای تشکیل X[k] و X[k+N/2].

در نتیجه در هر مرحله، تعداد عملیات نصف می‌شود و تعداد مراحل $log_2 N$ است، بنابراین کل پیچیدگی محاسباتی $O(N.log(N))$ می شود.

III. **محاسبه طیف توان (Power Spectrum) :** توان فرکانسی محاسبه می‌شود تا شدت هر فرکانس مشخص شود

$$P[k] = \frac{|X[k]|^2}{N}$$

برای محاسبه توان فرکانس k-ام، مربع دامنه ضرایب فرکانسی را تقسیم بر تعداد نمونه ها میکنیم.

نتایج:

ویژگی‌های استخراج‌شده از FFT در تحلیل صوتی نهایتا همچین چیزهایی هستند:

- فرکانس‌های غالب: شناسایی فرکانس‌هایی که بیشترین انرژی را دارند.
- طیف توان: توزیع انرژی سیگنال در دامنه فرکانس.

- تحلیل دامنه: بررسی تغییرات شدت صدا در بازه‌های زمانی مختلف.

از طرفی FFT محدودیت هایی دارد، مثلا:

- عدم ارائه اطلاعات زمانی: FFT اطلاعاتی درباره تغییرات سیگنال در زمان ارائه نمی‌دهد. این محدودیت با استفاده از Short-Time Fourier Transform رفع می‌شود.
- انتخاب طول فریم: طول فریم تاثیر مستقیمی بر دقت تحلیل دارد و فریم‌های کوتاه دقت زمانی بیشتر، اما رزولوشن فرکانسی کمتر می دهند و فریم‌های بلند دقت فرکانسی بیشتر، اما رزولوشن زمانی کمتری دارند.

نهایتا کاربردهای FFT در پردازش صوت را میتوان اینگونه قلمداد کرد:

- تشخیص گفتار: شناسایی فرکانس‌های پایه‌ای در گفتار انسان.
- تحلیل موسیقی: استخراج گام موسیقی و شناسایی سازها.
- کاهش نویز: شناسایی و حذف نویز فرکانسی از سیگنال صوتی.
- ویژگی‌سازی: ساخت ویژگی‌های پیشرفته مانند MFCC و Spectrogram که در ادامه راجع به آنها صحبت میکنیم.

2. Log Mel Spectrogram

توضیح اولیه و کلی

Log Mel Spectrogram به‌عنوان یک ویژگی کلیدی در پردازش سیگنال‌های صوتی و شناسایی الگوها، نقش حیاتی در استخراج اطلاعات معنادار از صوت ایفا می‌کند. این ویژگی یک نمایش فشرده و بهینه‌شده از سیگنال صوتی است که اطلاعات زمانی و فرکانسی آن را به شکلی سازمان‌یافته و قابل فهم برای مدل‌های یادگیری ماشین ارائه می‌دهد. Log Mel Spectrogram با تمرکز بر جنبه‌های مهم شنیداری و حذف داده‌های غیرضروری، به عنوان ابزاری قدرتمند برای تحلیل و پردازش صوتی شناخته می‌شود.

توضیح گام به گام و دقیق استخراج این ویژگی

I. در ابتدا دقیقا میتوانیم همان سه مرحله‌ی روش قبلی یعنی Fast Fourier Transform را اعمال کنیم که شامل تقسیم سیگنال به فریم های کوچک و هموار سازی آنها و سپس گرفتن FFT سیگنال و نهایتا محاسبه‌ی Power Spectrum می باشد.

II. **نگاشت فرکانس به مقیاس مل (Mel scale):** مقیاس مل مبتنی بر درک شنوایی انسان است که فرکانس‌های پایین را دقیق‌تر از فرکانس‌های بالا تشخیص می‌دهد. رابطه تبدیل فرکانس هرتز به مقیاس مل به صورت زیر است:

$$f_{\text{mel}} = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right)$$

III. **فیلتر بانک مل (Mel Filter Bank) :** سیگنال فرکانسی حاصل از FFT از مجموعه‌ای از فیلترهای مثلثی عبور داده می‌شود. این فیلترها در مقیاس مل تعریف شده‌اند تا انرژی فرکانس‌های مختلف را وزن‌دهی کنند.

برای هر فیلتر مثلثی $H_{m(f)}$ داریم:

$$H_m(f) = \begin{cases} 0 & f < f_{m-1} \\ \frac{f - f_{m-1}}{f_m - f_{m-1}} & f_{m-1} \le f < f_m \\ \frac{f_{m+1} - f}{f_{m+1} - f_m} & f_m \le f < f_{m+1} \\ 0 & f \ge f_{m+1} \end{cases}$$

که در آن f_{m-1}, f_m, f_{m+1} فرکانس‌های شروع، اوج، و پایان هر فیلتر بوده و $H_{m(f)}$ پاسخ فیلتر m-ام است.

IV. **برای هر فیلتر، انرژی محاسبه می شود:**

$$S_{\text{mel}}[m] = \sum_{k} P[k] \cdot H_m[k]$$

در این فرمول ها، $H_m[k]$ وزن فیلتر m-ام برای فرکانس k است و $P[k]$ هم همان Power Spectrum یا توان فرکانس k-ام است که در روش قبلی (FFT) به تفصیل روابط آن بیان شد. و نتیجه $S_{mel}[m]$ انرژی مقیاس مل برای فیلتر m-ام است.

۷. **گرفتن لگاریتم:** برای فشرده‌سازی دامنه‌ها و تطبیق با سیستم شنوایی انسان (درک لگاریتمی)، لگاریتم انرژی‌های مل گرفته می‌شود:

$$\text{Log Mel Spectrogram}[m, t] = \log(S_{\text{mel}}[m, t] + \epsilon)$$

که در آن ϵ یک مقدار کوچک برای جلوگیری از صفر شدن در لگاریتم است.

نتایج

با ترکیب تمام مراحل، Log Mel Spectrogram برای یک سیگنال صوتی مثل x[n] به صورت زیر است:

$$\text{Log Mel Spectrogram}[m, t] = \log\left(\sum_{k} \frac{|X_t[k]|^2}{N} \cdot H_m[k] + \epsilon\right)$$

نهایتا طیف‌نگار لگاریتمی مل معمولاً به صورت یک تصویر دوبعدی نمایش داده می‌شو که در آن محور x: زمان (فریم‌ها) است و محور فرکانس (در مقیاس مل) بوده و شدت رنگ هم دامنه بصورت لگاریتمی است.

کاربردهای این روش در پردازش صوت را میتوان اینگونه قلمداد کرد:

- تشخیص گفتار (Speech Recognition) به عنوان ورودی به شبکه‌های عصبی.
- تحلیل موسیقی مثل شناسایی گام، آکورد، و نوع موسیقی.
- دسته‌بندی صداهای محیطی یعنی مثلا تمایز بین انواع صداهای طبیعی و مصنوعی.

3. Mel Frequency Cepstral Coefficients (MFCC)

توضیح اولیه و کلی

MFCC یکی از پرکاربردترین تکنیک‌های استخراج ویژگی در پردازش صوت، به‌ویژه در تشخیص گفتار، پردازش موسیقی، و تحلیل صوت است. این روش سیگنال صوتی را از دامنه زمانی به دامنه فرکانسی و سپس به ضرایب **Cepstral** در مقیاس فرکانس مل نگاشت می‌کند. MFCC با استفاده از تبدیل‌های ریاضی دقیق، ویژگی‌هایی را استخراج می‌کند که بهتر با سیستم شنوایی انسان تطبیق دارند.

توضیح گام به گام و دقیق استخراج این ویژگی

I. در ابتدا دقیقا میتوانیم همان روش قبلی یعنی Log Mel Spectrogram را اعمال کنیم که شامل تقسیم سیگنال به فریم های کوچک و هموار سازی آنها و سپس گرفتن FFT سیگنال و و محاسبه‌ی Power Spectrum و بعد از آن نگاشت به مقیاس مل و اعمال بانک فیلتر مل و محاسبه ی لگاریتم انرژی است.

II. **تبدیل کسینوسی گسسته (DCT) :** برای فشرده‌سازی اطلاعات و کاهش همبستگی بین ضرایب، تبدیل کسینوسی گسسته (DCT) اعمال می‌شود (فقط ضرایب پایین‌تر از مثلا n=13 نگهداری می‌شوند، زیرا اطلاعات اصلی در آن‌ها متمرکز است.) [2]

$$c_n = \sum_{m=1}^{M} E_m^{\log} \cdot \cos\left[\frac{\pi n}{M}\left(m - \frac{1}{2}\right)\right]$$

که در آن M تعداد فیلترهای بانک مل بوده و c_nها ضرایب MFCC نام دارند.

III. **Delta and Delta-Delta Coefficients :** برای در نظر گرفتن پویایی‌های زمانی، مشتقات اول و دوم ضرایب کپسترال نیز محاسبه می‌شوند:

$$\Delta c_n[t] = \frac{\sum_{k=1}^{K} k \cdot (c_n[t+k] - c_n[t-k])}{2 \cdot \sum_{k=1}^{K} k^2}$$

که در آن $\Delta c_n[t]$ ها مشتق اول c_n در زمان t هستند و K اندازه پنجره برای مشتق‌گیری است.

IV. **نتایج:** با تلفیق تمام مراحل قبلی، داریم:

$$c_n = \sum_{m=1}^{M} \log \left(\sum_{k=1}^{K} \frac{|X[k]|^2}{N} \cdot H_m[k] \right) \cdot \cos \left[\frac{\pi n}{M} \left(m - \frac{1}{2} \right) \right]$$

پس نهایتا به عنوان ویژگی‌های MFCC داریم:

- ضرایب $c_1, c_2, \ldots$ ویژگی‌های اصلی هستند.
- Delta MFCC نرخ تغییر ضرایب MFCC هستند.
- Delta-Delta MFCC: شتاب تغییرات ضرایب هستند.

کاربردهای MFCC را هم میتوان اینگونه بیان کرد:

- تشخیص گفتار: ورودی به سیستم‌های شناسایی خودکار گفتار (ASR).
- تحلیل موسیقی: شناسایی ژانر و تشخیص سازها.
- شناسایی گوینده: استخراج ویژگی‌های منحصر به فرد هر گوینده.

این روش مزایای زیادی دارد مثلا اینکه همانند روش قبلی، با استفاده از مقیاس مل، MFCC شبیه به نحوه شنیدن انسان عمل می‌کند، همچنین کاهش ابعاد باعث می شود ویژگی‌های ضروری در یک نمایش فشرده نگه داشته می‌شوند و در نهای کاربرد گسترده ای داشته و قابل استفاده در انواع تحلیل‌های صوتی است.

4. Spectral Centroid

توضیح اولیه و کلی

Spectral Centroid یک ویژگی پرکاربرد در پردازش صوت است که به عنوان "مرکز ثقل" طیف تعریف می‌شود. این ویژگی نشان می‌دهد که میانگین وزنی فرکانس‌های موجود در سیگنال کجا قرار دارد. این ویژگی با شدت فرکانس‌ها وزن‌دهی می‌شود و به درک "روشنایی" یا "تیزی" صدا مرتبط است.

توضیح گام به گام و دقیق استخراج این ویژگی

I. در ابتدا میتوان از مراحل یافت FFT کرد که قبلا هم بیان شدند، یعنی سیگنال صوتی به بازه‌های زمانی کوتاه (TT) تقسیم شود تا ویژگی‌های زمان-محلی طیف بررسی شوند و برای کاهش ناپیوستگی در لبه‌های فریم، هر فریم با یک تابع پنجره ضرب شود و با استفاده از FFT، سیگنال از دامنه زمان به دامنه فرکانس تبدیل شود.

II. **محاسبه ی Magnitude Spectrum یا بزرگی طیف :** بزرگی طیف (S[k]) برابر مقدار مطلق FFT است:

$$S[k] = |X[k]|$$

III. **Map Frequencies to FFT Bins :** فرکانس متناظر با هر باند k به صورت زیر محاسبه می‌شود:

$$f_k = \frac{k \cdot f_s}{N}$$

که در آن f_s نرخ نمونه‌برداری سیگنال صوتی بوده و N تعداد نقاط FFT است.

IV. محاسبه‌ی Spectral Centroid یا مرکز ثقل طیفی : با استفاده از فرمول میانگین وزنی داریم [3]

$$C = \frac{\sum_{k=1}^{N} f_k \cdot S[k]}{\sum_{k=1}^{N} S[k]}$$

همانطور که گفتیم f_k فرکانس متناظر با باند k است و S[k] برابر با بزرگی طیف در باند k است و N تعداد باندهای فرکانسی است.

تفسیر فیزیکی این ویژگی به این صورت است که مرکز ثقل طیفی بالا نشان‌دهنده این است که انرژی بیشتر در فرکانس‌های بالا متمرکز است. این معمولاً در صداهای تیز و روشن (مانند صدای سنج) مشاهده می‌شود و بالعکس، مرکز ثقل طیفی پایین نشان‌دهنده متمرکز بودن انرژی در فرکانس‌های پایین است. این ویژگی در صداهای تاریک و نرم (مانند گیتار باس) قابل مشاهده است.

نتایج:

استفاده ازین ویژگی مزایایی دارد مثلا:

- این ویژگی با درک انسان از تن صدا و روشنایی صدا مرتبط است.
- محاسبه مرکز ثقل طیفی از طریق FFT ساده و سریع است.
- این ویژگی در تحلیل گفتار، موسیقی، و صداهای محیطی مفید است.

کاربردهای این روش عبارتند از:

- تحلیل تن صدا و مثلا شناسایی روشنایی یا تیزی در سازهای موسیقی.
- کارهای مربوط به پردازش گفتار مثل تمایز بین واج‌ها یا ویژگی‌های گوینده.
- رده‌بندی ژانر موسیقی و تفکیک ژانرها بر اساس ویژگی‌های تیزی صدا.
- شناسایی محتوای طیفی صداهای محیطی.

اما استفاده از این روش در عمل یکسری مشکلاتی دارد مثلا:

- حساسیت به نویز: مرکز ثقل طیفی ممکن است به نویز در فرکانس‌های بالا حساس باشد پس استفاده از فیلترهای حذف نویز ضروری است.
- وابستگی به طول فریم: انتخاب طول مناسب فریم اهمیت زیادی دارد و فریم‌های کوتاه باعث وضوح زمانی بهتر(دقت در تشخیص تغییرات سریع در سیگنال) اما جزئیات کمتر در فرکانس می شوند و فریم‌های بلند منجر به وضوح فرکانسی بهتر (توانایی در تمیز دادن فرکانس‌های نزدیک به هم از یکدیگر) اما جزئیات کمتر در زمان می شوند.
- پیچیدگی طیفی: سیگنال‌های پیچیده با چندین قله طیفی ممکن است به توصیف‌های اضافی نیاز داشته باشند.
- این مقاله [4] هم برای شناخت سایر ویژگی های داده های صوتی و هم همین ویژگی جاری، به خوبی و دقت و جزئیات توضیح داده است اما مقصود این پیش گزارش، دقتی در این حد نیست، پس به بیان آن بسنده می کنیم.

5. Chroma Features

توضیح اولیه و کلی

ویژگی‌های کروماتیک (Chroma Features) که با نام بردارهای کروماتیک (Chroma Vectors) نیز شناخته می‌شوند، توزیع انرژی یا شدت سیگنال صوتی را در ۱۲ کلاس زیروبمی کروماتیک (C، C#، D،

B ...) بدون توجه به اکتاو نشان می‌دهند. این ویژگی‌ها به‌طور گسترده در حوزه تحلیل اطلاعات موسیقی برای شناسایی آکورد، تشخیص گام، و طبقه‌بندی ژانر موسیقی استفاده می‌شوند. [5]

پس ویژگی‌های کروماتیک، مشخصه‌های هارمونیک و ملودیک سیگنال صوتی را با گروه‌بندی تمام فرکانس‌های مرتبط با یک کلاس زیروبمی یکسان (مانند C در هر اکتاو) در یک ویژگی ارائه می‌دهند..نکته ی مهم در این ویژگی ها، بی توجهی به اکتاو است یعنی فرکانس‌هایی که تنها به دلیل تفاوت در اکتاو متفاوت هستند (مانند ۲۶۱.۶۳ هرتز برای C میانی و ۵۲۳.۲۵ هرتز برای C بعدی) در یک کلاس قرار می‌گیرند.

چرا بی‌توجهی به اکتاو مفید است؟

- سادگی تحلیل: بی‌توجهی به اکتاو، ابعاد نمایش ویژگی‌ها را کاهش می‌دهد و فقط ۱۲ ویژگی (برای ۱۲ کلاس کروماتیک) باقی می‌ماند.این کاهش ابعاد باعث تسهیل در تحلیل و پردازش داده‌ها می‌شود.
- هارمونی موسیقی: در موسیقی، اکتاوها معمولاً مکمل هم هستند و اطلاعات زیادی در تفاوت‌های بین اکتاوها وجود ندارد. برای مثال، آکوردهای موسیقی از کلاس‌های زیروبمی مشخصی تشکیل شده‌اند و اکتاو تأثیر زیادی بر هارمونی ندارد.

- پیش از بیان مراحل، لازم است مقدمه ای بر تفاوت FFT و STFT بیان شود چون ما تا به اینجا از STFT نام نبردیم اما در اصل با عملیات فریم بندی و اعمال FFT داشتیم STFT میگرفتیم.

تبدیل فوریه کوتاه‌مدت (STFT) : همانطور که گفتیم برای هر فریم، تبدیل فوریه کوتاه‌مدت محاسبه می‌شود تا نمایش فرکانسی سیگنال به دست آید:

$$X[k,t] = \sum_{n=0}^{N-1} x_w[n] \cdot e^{-j\frac{2\pi kn}{N}}$$

که در آن X[k, t] ضرایب فرکانسی مختلط برای فریم t و باند فرکانسی k را نشان می دهد و N تعداد نقاط FFT است

۱. تفاوت بین FFT و STFT

:(FFT) Fast Fourier Transform

- تبدیل کل سیگنال به دامنه فرکانس بدون در نظر گرفتن زمان.
- نتیجه FFT برای کل سیگنال یک طیف فرکانسی کلی است که نشان می‌دهد چه فرکانس‌هایی در کل سیگنال وجود دارند، اما هیچ اطلاعاتی درباره زمان وقوع آن‌ها ارائه نمی‌دهد.

:(STFT) Short-Time Fourier Transform

- تقسیم سیگنال به فریم‌های کوتاه و اعمال FFT به هر فریم.
- نتیجه STFT یک ماتریس است که هر ستون آن نشان‌دهنده طیف فرکانسی یک بازه زمانی خاص است.
- اطلاعات فرکانسی به صورت زمان-محلی (Time-Frequency) ارائه می‌شود.

در نتیجه STFT برای ویژگی‌های کروماتیک استفاده می‌شود چون سیگنال‌های موسیقی و گفتار معمولاً پویا هستند و فرکانس‌های آن‌ها در طول زمان تغییر می‌کند. و ویژگی‌های کروماتیک باید نشان دهند که در هر لحظه چه فرکانس‌هایی (کلاس‌های زیروبمی) غالب هستند و با STFT، می‌توان توزیع انرژی در کلاس‌های زیروبمی را به صورت فریم به فریم محاسبه کرد چون STFT سیگنال را به فریم‌های کوتاه تقسیم می‌کند و اجازه می‌دهد هر فریم جداگانه تحلیل شود. پس STFT این امکان را می‌دهد که تغییرات

دینامیکی سیگنال در طول زمان (مانند تغییر گام یا آکورد) به‌خوبی ثبت شوند. نهایتا FFT معمولی برای سیگنال‌های ایستا مناسب است، اما در سیگنال‌های پویا اطلاعات زمانی را از دست می‌دهد.

توضیح گام به گام و دقیق

I. در مرحله‌ی اول میتوان از اعمالی که در اولین مورد که FFT بود ذکر کردیم استفاده شود، همانند پیش پردازش و تقسیم به فریم ها و اعمال پنجره، البته گفتیم که نهایتا گرفتن FFT چه در این روش و چه در روش های دیگر چون فریم به فریم اعمال میشد در اصل تبدیل فوریه ی کوتاه مدت یا STFT است که در همین بخش توضیح کلی ای بر آن دادیم. نهایتا هم در همان روش طیف بزرگی محاسبه میشد که با S[k] نشان می دادیم آن را.

II. **نگاشت فرکانس‌ها به کلاس‌های زیروبمی (Pitch Classes):** فرکانس‌های موجود در طیف به ۱۲ کلاس کروماتیک زیروبمی نگاشت می‌شوند که اولین مرحله ی آن محاسبه اندیس کلاس زیروبمی (Pitch Class Index):

$$p = \left(\text{round} \left(12 \cdot \log_2 \left(\frac{f_k}{f_{\text{ref}}} \right) \right) \mod 12 \right)$$

در عبارت بالا، f_k فرکانس متناظر با باند k است و f_{ref} فرکانس مرجع یا همان A در اکتاو چهارم (A4) با مقدار 440 هرتز است و p اندیس کلاس کروماتیک است صفر برای C، یک برای C# و ... و 11 برای B.

III. **جمع انرژی طیفی برای هر کلاس زیروبمی:**

$$C[p] = \sum_{f_k \in \text{octaves}} S[k] \cdot \delta(f_k, p)$$

$\delta(f_k, p)$: تابع شاخص که تعیین می‌کند f_k به کلاس کروماتیک p تعلق دارد یا خیر.

یا به بیان ساده تر:

$$C[p] = \sum_{k:\text{pitch}(f_k)=p} S[k]$$

در کل این عمل محتوای هارمونی تمام اکتاوها را در یک کلاس جمع میکند.

IV. **نرمال‌سازی :** ویژگی‌های کروماتیک نرمال‌سازی می‌شوند تا اثر تفاوت‌های شدت سیگنال حذف شود:

$$C_{\text{norm}}[p] = \frac{C[p]}{\sum_{q=0}^{11} C[q]}$$

این باعث می شود که مجموع تمام chroma features مساوی 1 شود.

نتایج

فرمول کلی برای محاسبه ویژگی‌های کروماتیک در هر فریم به صورت زیر است:

$$C_{\text{norm}}[p] = \frac{\sum_{k:\text{pitch}(f_k)=p} |X[k]|}{\sum_{q=0}^{11} \sum_{k:\text{pitch}(f_k)=q} |X[k]|}$$

کاربردهای این ویژگی عبارتند از:

- تعیین آکوردها بر اساس کلاس‌های کروماتیک غالب.

- تشخیص گام موسیقی، یعنی تحلیل کلی توزیع کلاس‌های کروماتیک برای یافتن گام موسیقی.
- طبقه‌بندی ژانر موسیقی و تمایز ژانرها با تحلیل محتوای هارمونیک.
- شناسایی آهنگ‌های مشابه (مقایسه ویژگی‌های کروماتیک برای یافتن شباهت ساختاری)

مزایای استفاده از Chroma Features هم همانطور که گفتیم، شامل بی‌توجهی به اکتاو، سادگی محاسبات(با استفاده از FFT و نگاشت ساده، ویژگی‌های کروماتیک به راحتی محاسبه می‌شوند.)، و این نکته است این ویژگی‌ها با ساختار هارمونیک موسیقی غربی کاملاً سازگار هستند.

چالش‌ها

- موسیقی غیر غربی: ویژگی‌های کروماتیک برای موسیقی‌هایی که از مقیاس ۱۲ نیم‌پرده‌ای پیروی نمی‌کنند کمتر کاربرد دارند.
- حساسیت به نویز: نویز ممکن است بر دقت نگاشت فرکانس‌ها به کلاس‌های کروماتیک تاثیر بگذارد.
- ابهام در هارمونیک‌ها: هم‌پوشانی فرکانس‌های هارمونیک ممکن است به دقت ویژگی‌ها آسیب بزند.

6. Spectral Contrast

توضیح اولیه و کلی

Spectral Contrast یکی از ویژگی‌های مهم در تحلیل صوت است که به تفاوت انرژی بین فرکانس‌های بالا و پایین در یک بازه زمانی مشخص تمرکز دارد. این ویژگی اطلاعات مفیدی درباره ساختار هارمونیکی و نویزی سیگنال ارائه می‌دهد. Spectral Contrast در کاربردهایی مانند تشخیص ژانر موسیقی، طبقه‌بندی صوت، و شناسایی گفتار استفاده می‌شود. این روش به خصوص برای تمایز بین صداهای موسیقایی و غیر موسیقایی یا صداهای پیچیده و ساده بسیار کارآمد است.

توضیح گام به گام و دقیق استخراج این ویژگی

I. **تقسیم سیگنال به فریم‌های کوچک و هموارسازی آنها:**

در گام اول، سیگنال صوتی به فریم‌های کوچک تقسیم می‌شود (مانند Windowing) و سپس یک پنجره مناسب (مانند پنجره Hamming) روی هر فریم اعمال می‌شود تا از اثرات ناگهانی جلوگیری شود.

II. **محاسبه FFT:**

برای تبدیل سیگنال از دامنه زمانی به دامنه فرکانسی، FFT روی هر فریم اعمال می‌شود. این تبدیل طیف فرکانسی سیگنال را فراهم می‌کند.

III. **تقسیم طیف به بازه‌های فرکانسی (Subbands):**

طیف فرکانسی به چندین باند فرکانسی تقسیم می‌شود. تعداد این باندها (مثلاً 6 یا 7) بستگی به کاربرد دارد. این باندها معمولاً به صورت لگاریتمی توزیع می‌شوند تا فرکانس‌های پایین‌تر با دقت بیشتری تحلیل شوند.

IV. **محاسبه بیشینه و کمینه انرژی در هر باند:**

برای هر باند فرکانسی، مقادیر بیشینه (Peaks: نشان‌دهنده قوی‌ترین فرکانس‌های موجود در هر باند) و کمینه (Valleys: مشخص کننده انرژی ضعیف‌ترین نقاط) انرژی محاسبه می‌شوند.

V. **محاسبه نسبت کنتراست طیفی (Spectral Contrast):**

کنتراست طیفی به صورت تفاوت لگاریتمی بین Peaks و Valleys در هر باند تعریف می‌شود:

$$SpectralContrast(i) = log(PeakEnergy(i)ValleyEnergy(i))\text{Spectral Contrast}(i) = \log\left(\frac{\text{Peak Energy}(i)}{\text{Valley Energy}(i)}\right)$$

که در آن i شاخص باند فرکانسی است.

VI. **انجام میانگین‌گیری در فریم‌ها:**

مقادیر Spectral Contrast به طور میانگین در کل فریم‌ها محاسبه می‌شوند تا یک نمای کلی از کنتراست طیفی سیگنال ایجاد شود.

نتایج

با جمع‌آوری این مراحل، مقادیر نهایی Spectral Contrast برای سیگنال صوتی به دست می‌آید. این ویژگی اطلاعاتی را درباره توزیع انرژی بین فرکانس‌های مختلف ارائه می‌دهد، که برای تحلیل و طبقه‌بندی صوت به باندهای فرکانسی با انرژی بالا (که معمولا در صداهای موسیقیایی با هارمونی بالا دیده می‌شوند) باندهای فرکانسی با انرژی پایین (که مشخص‌کننده سیگنال‌های نویزی یا صداهای ساده‌تر هستند) بسیار مفید است.

کاربردهای Spectral Contrast

از کاربردهای Spectral Contrast میتوان به تشخیص ژانر موسیقی [10] اشاره کرد که در آن می‌توان ژانرهایی مانند موسیقی با کنتراست بالا (مانند موسیقی کلاسیک) و موسیقی با کنتراست پایین (مانند موسیقی الکترونیک) را از یکدیگر تمایز داد. همچنین، این روش برای شناسایی گوینده نیز به کار می‌رود، زیرا می‌توان ویژگی‌هایی را استخراج کرد که الگوی خاص هر گوینده را مشخص کنند. از دیگر کاربردهای آن، تحلیل محیط صوتی است که به تشخیص تفاوت بین صداهای محیطی و موسیقیایی کمک می‌کند.

مزایا

I. **کاربرد گسترده:** Spectral Contrast می‌تواند در طیف وسیعی از کاربردهای صوتی، از جمله شناسایی موسیقی و صدا، استفاده شود.

II. **حساسیت بالا به ساختار هارمونیکی:** این ویژگی اطلاعات دقیق‌تری درباره ساختار طیفی صدا نسبت به ویژگی‌های ساده‌تری مانند Spectral Centroid ارائه می‌دهد.

III. **تحلیل قابل انعطاف:** امکان تنظیم تعداد باندهای فرکانسی برای تطبیق با کاربردهای مختلف.

معایب

I. **حساسیت به نویز:** Spectral Contrast ممکن است در محیط‌های پر از نویز دقت کمتری داشته باشد.

II. **نیاز به تنظیم دقیق:** انتخاب تعداد باندهای فرکانسی و پارامترهای دیگر نیازمند تنظیمات دقیق است تا نتایج مطلوب حاصل شود.

III. **پیچیدگی محاسباتی:** محاسبه Spectral Contrast نسبت به ویژگی‌های ساده‌تر به منابع محاسباتی بیشتری نیاز دارد.

7. Zero-Crossing Rate (ZCR)

توضیح اولیه و کلی

Zero-Crossing Rate (ZCR) یکی از ساده‌ترین و موثرترین ویژگی‌ها در پردازش صوت است که تعداد دفعات عبور سیگنال از محور صفر را در یک بازه زمانی مشخص اندازه‌گیری می‌کند. این ویژگی معمولاً برای تحلیل ویژگی‌های زمانی سیگنال و شناسایی انواع صداها مانند صداهای گفتاری، موسیقایی، یا نویزی استفاده می‌شود[10]. ZCR به ویژه در تشخیص بی‌صدا (unvoiced) و صدادار (voiced) بودن گفتار کاربرد دارد، زیرا نرخ عبور صفر در صداهای بی‌صدا معمولاً بالاتر است.

توضیح گام به گام و دقیق محاسبه ZCR

I. تقسیم سیگنال به فریم‌های کوچک:

ابتدا سیگنال صوتی به بخش‌های کوچک (فریم‌ها) تقسیم می‌شود. این تقسیم‌بندی معمولاً با پنجره‌گذاری (مانند Hamming یا Rectangular) انجام می‌شود. اندازه فریم به کاربرد بستگی دارد (مثلاً 20 تا 40 میلی‌ثانیه).

II. محاسبه تعداد عبور از صفر برای هر فریم:

Zero-crossing زمانی رخ می‌دهد که علامت (sign) سیگنال از مثبت به منفی یا برعکس تغییر کند. این تغییرات به صورت زیر فرمول‌بندی می‌شوند:

$$ZCR = \frac{1}{N-1} \sum_{n=1}^{N-1} \mathbf{1}\left(s[n] \cdot s[n-1] < 0\right)$$

که در آن N تعداد نمونه‌ها در فریم، s[n] مقدار نمونه در زمان n و 1(x) تابع نشانگر که مقدار 1 را برمی‌گرداند اگر شرط x صحیح باشد وگرنه 0.

III. نرمال‌سازی مقدار ZCR:

مقدار ZCR معمولاً بر اساس تعداد نمونه‌ها نرمال‌سازی می‌شود تا مستقل از طول فریم باشد. این نرمال‌سازی باعث می‌شود ویژگی برای مقایسه بین فریم‌ها استاندارد شود.

نتایج

Zero-Crossing Rate برای هر فریم محاسبه می‌شود و به عنوان یک مقدار سریالی یا میانگین کلی برای کل سیگنال گزارش می‌شود. مقادیر بالای ZCR معمولاً در صداهای صامت (مانند صدای "s" یا نویزهای سفید) و مقادیر پایین آن در در صداهای مصوت دیده می‌شود که نوسانات کمتری دارند.

کاربردهای Zero-Crossing Rate

از کاربردهای Zero-Crossing Rate میتوان به تشخیص گفتار و موسیقی برای تمایز بین بخش‌های گفتاری و موسیقیایی در یک فایل صوتی، شناسایی گوینده و گفتار در سیستم‌های تشخیص گفتار، طبقه‌بندی صوتی برای شناسایی نویزها و صداهای با فرکانس بالا یا پایین و آنالیز ریتم موسیقی به عنوان ویژگی‌ای برای تشخیص و شناسایی الگوهای ریتمیک اشاره کرد.

مزایا

I. **سادگی محاسبه:** ZCR به محاسبات پیچیده‌ای نیاز ندارد و به راحتی قابل استخراج است.
II. **تشخیص سریع صدا:** نرخ عبور از صفر اطلاعات اولیه مفیدی درباره ماهیت سیگنال فراهم می‌کند.
III. **استقلال از دامنه سیگنال:** ZCR به دامنه سیگنال حساس نیست و به تغییرات قدرت (amplitude) وابسته نیست.

معایب

I. **حساسیت به نویز:** نویز می‌تواند نرخ عبور از صفر را افزایش داده و نتایج را تحریف کند.
II. **محدودیت در کاربردهای فرکانسی پایین:** در تحلیل صداهای پیچیده‌تر، ممکن است ZCR کافی نباشد.

8. Linear Predictive Coding (LPC)

توضیح اولیه و کلی

Linear Predictive Coding (LPC) یکی از روش‌های برجسته در پردازش صوت است که ویژگی‌های مهم صوتی را با مدل‌سازی طیف سیگنال به صورت خطی استخراج می‌کند. این روش بر این فرض تکیه دارد که هر نمونه صوتی را می‌توان به صورت ترکیبی خطی از مقادیر گذشته پیش‌بینی کرد. LPC به طور گسترده در تحلیل گفتار، شناسایی گوینده، و فشرده‌سازی صوت استفاده می‌شود[11].

هدف اصلی LPC استخراج ضرایب خطی‌ای است که بهترین تقریب را از سیگنال صوتی ارائه می‌دهند، و این ضرایب اطلاعات مهمی درباره ساختار گفتار، مانند فرکانس‌های فرمات، در اختیار قرار می‌دهند.

توضیح گام به گام و دقیق محاسبه LPC

I. تقسیم سیگنال به فریم‌های کوچک:

برای تحلیل سیگنال، ابتدا آن را به فریم‌های کوچک تقسیم می‌کنیم (معمولاً بین 20 تا 40 میلی‌ثانیه) تا سیگنال در هر فریم تقریبی از یک سیگنال ایستا (stationary) باشد.

II. محاسبه‌ی تابع خودهمبستگی (Autocorrelation):

تابع خودهمبستگی سیگنال برای هر فریم محاسبه می‌شود:

$$R(k) = \sum_{n=k}^{N-1} x(n)x(n-k)$$

که در آن N تعداد نمونه‌ها در فریم، R(k) مقدار خود همبستگی برای تاخیر k و x(n) مقدار نمونه در زمان n میباشد.

III. حل معادلات خطی (Levinson-Durbin Algorithm):

ضرایب پیش‌بینی خطی با استفاده از تابع خودهمبستگی و الگوریتم Levinson-Durbin محاسبه می‌شوند:

$$R(k) = \sum_{i=1}^{p} a_i R(k-i)$$

که در آن p مرتبه مدل LPC است. این الگوریتم با کمترین پیچیدگی محاسباتی، ضرایب ai را به دست می‌آورد.

IV. محاسبه خطای پیش‌بینی (Prediction Error):

خطای پیش‌بینی یا Residual Signal از تفاضل سیگنال واقعی و سیگنال پیش‌بینی‌شده محاسبه می‌شود:

$$e(n) = x(n) - \sum_{i=1}^{p} a_i x(n-i)$$

این سیگنال معمولاً شامل صداهای ناگهانی (excitation signal) مانند انفجارهای صوتی یا نویز در گفتار است.

V. تخمین پارامترهای طیفی (Spectral Parameters):

ضرایب LPC معمولاً برای استخراج ویژگی‌های طیفی، مانند فرکانس‌های فرمات، مورد استفاده قرار می‌گیرند. طیف LPC یک تخمین صاف از پاسخ فرکانسی سیگنال است.

نتایج

در پایان مراحل فوق، ضرایب a1,a2,..., ap به عنوان ضرایب پیش‌بینی خطی محاسبه می‌شوند که نشان‌دهنده ویژگی‌های کلیدی و مهم سیگنال صوتی می‌باشند. علاوه بر این، سیگنال Residual اطلاعات مرتبط با منبع سیگنال، مانند شدت و نرخ ضربان، را در خود حفظ می‌کند. طیف LPC نیز به عنوان یک نمایش طیفی صاف از سیگنال عمل می‌کند که می‌تواند برای استخراج ویژگی‌های طیفی دقیق و کاهش نویز مورد استفاده قرار گیرد. این ترکیب از ضرایب و سیگنال Residual ابزار قدرتمندی برای تحلیل و پردازش سیگنال‌های صوتی فراهم می‌آورد.

کاربردهای LPC

کاربردهای LPC شامل تحلیل گفتار، شناسایی گوینده (به طوری که ضرایب LPC می‌توانند به عنوان ویژگی‌های منحصر به فرد برای تشخیص گوینده استفاده شوند)، فشرده‌سازی صوت (در کدک‌های صوتی مانند GSM از LPC برای کاهش حجم داده استفاده می‌شود) و تبدیل گفتار به متن (ASR) می‌باشند.

مزایا

I. **مدل‌سازی دقیق طیف:** LPC می‌تواند ساختار طیفی سیگنال را به طور موثر مدل کند.
II. **استخراج ویژگی‌های مهم:** ضرایب LPC اطلاعاتی مانند فرکانس‌های فرمات را استخراج می‌کنند که برای کاربردهای گفتار و صوت حیاتی هستند.
III. **کاربرد گسترده:** LPC در تحلیل، شناسایی، و فشرده‌سازی صوت استفاده می‌شود.

معایب

I. **حساسیت به نویز:** در محیط‌های نویزی، کارایی LPC ممکن است کاهش یابد.

II. **محدودیت در مدل‌سازی سیگنال‌های پیچیده:** سیگنال‌هایی با ویژگی‌های غیر خطی یا غیر ایستا ممکن است به درستی مدل نشوند.

9. Perceptual Linear Prediction (PLP)

توضیح اولیه و کلی

Perceptual Linear Prediction (PLP) یک روش پیشرفته در پردازش گفتار است که از اصول روان‌آکوستیکی برای بهبود دقت مدل‌سازی طیفی استفاده می‌کند[11]. این روش در واقع اصلاحی از LPC است که با در نظر گرفتن ویژگی‌های شنوایی انسان، مانند حساسیت فرکانسی و اثرات ماسکینگ، به بازنمایی دقیق‌تر سیگنال گفتاری می‌پردازد. PLP به طور گسترده در سیستم‌های شناسایی گفتار و گفتار به متن (ASR) به کار گرفته می‌شود.

هدف اصلی PLP، کاهش پیچیدگی غیرضروری در طیف سیگنال با استفاده از اصول شنوایی است، به‌طوری‌که بازنمایی به دست آمده بهتر با نحوه شنیدن انسان تطابق داشته باشد.

توضیح گام به گام و دقیق محاسبه PLP

I. **محاسبه طیف توان سیگنال (Power Spectrum):**

ابتدا طیف توان سیگنال را تعریف میکنیم. این تابع ورودی را با استفاده از تبدیل فوریه (FFT) به حوزه فرکانس تبدیل کرده، و طیف توان آن را به شکل زیر محاسبه میکند:

$$P(f) = |FFT(f)|^2$$

II. اعمال مقیاس بارک (Bark Scale):

طیف توان به یک مقیاس فرکانسی غیرخطی (مقیاس بارک) نگاشت می‌شود. این مقیاس، حساسیت شنوایی انسان به فرکانس‌های مختلف را شبیه‌سازی می‌کند:

$$z(f) = 6 \cdot \ln\left(\frac{f}{600} + \sqrt{\left(\frac{f}{600}\right)^2 + 1}\right)$$

که z(f) فرکانس در مقیاس بارک است.

III. اعمال فیلتر روان‌آکوستیکی (Critical Band Filtering):

طیف بارک توسط مجموعه‌ای از فیلترهای گوسی که پهنای باند بحرانی را شبیه‌سازی می‌کنند، هموار می‌شود. این مرحله تأثیر ماسکینگ فرکانسی در سیستم شنوایی انسان را شبیه‌سازی می‌کند.

IV. اعمال تقلیل دینامیک طیفی (Spectral Compression):

برای بهبود مطابقت با سیستم شنوایی انسان، طیف هموار شده با یک تابع لگاریتمی یا توان فشرده می‌شود:

$$P'(z) = P(z)^{0.33}$$

این تابع تاثیر کاهش حساسیت گوش به تغییرات کوچک در شدت صوت را مدل می‌کند.

V. محاسبه ضرایب LPC:

طیف اصلاح‌شده P′(z) به حوزه زمان بازگشته و ضرایب LPC از آن استخراج می‌شود، مشابه روش سنتی LPC.

VI. تبدیل به ضرایب PLP:

ضرایب LPC محاسبه شده با وزن دهی فرکانسی و کاهش مرتبه بهینه سازی و به ضرایب PLP تبدیل می‌شوند. این ضرایب به عنوان ویژگی‌های نهایی مورد استفاده قرار می‌گیرند.

نتایج

در پایان این فرآیند ضرایب PLP ویژگی‌هایی هستند که اطلاعات مهم گفتار را در قالبی فشرده و مقاوم در برابر نویز ارائه می‌دهند. طیف بهینه‌شده نیز بازنمایی طیفی می‌باشد که بر اساس اصول شنوایی انسان تعدیل شده است.

کاربردهای PLP

کاربردهای PLP شامل شناسایی خودکار گفتار (ASR) است، جایی که ضرایب PLP به دلیل بازنمایی دقیق‌تر گفتار برای سیستم‌های تبدیل گفتار به متن مورد استفاده قرار می‌گیرند. همچنین، این روش در تحلیل گفتار برای استخراج ویژگی‌های مهم با دقت بالا، به‌ویژه در محیط‌های واقعی، کاربرد دارد. علاوه بر این، در فشرده‌سازی صوت نیز از PLP برای بهینه‌سازی داده‌های صوتی و کاهش پیچیدگی طیفی غیرضروری بهره گرفته می‌شود

مزایا

1. **مقاومت در برابر نویز:** با در نظر گرفتن اصول شنوایی، PLP به طور قابل توجهی نسبت به LPC به نویز کمتر حساس است.
2. **بازنمایی دقیق‌تر گفتار:** مدل‌سازی ویژگی‌های گفتاری بر اساس سیستم شنوایی انسان، دقت تحلیل را افزایش می‌دهد.
3. **کاربردهای گسترده:** PLP به طور گسترده در سیستم‌های ASR و تحلیل گفتار استفاده می‌شود.

معایب

1. **پیچیدگی محاسباتی بالاتر:** مراحل اضافی مانند اعمال مقیاس بارک و فیلترهای بحرانی، PLP را نسبت به LPC پیچیده‌تر می‌کند.
2. **حساسیت به تنظیم پارامترها:** کیفیت نتایج PLP به تنظیم دقیق پارامترهای مدل بستگی دارد.

Similarity Learning

در عصر حاضر، فناوری‌های صوتی به عنوان یکی از ابزارهای حیاتی در حوزه‌های مختلف از جمله امنیت، ارتباطات و خدمات دیجیتال شناخته شده‌اند. **Similarity Learning** یا **یادگیری شباهت** به عنوان یکی از شاخه‌های مهم یادگیری ماشین، نقش کلیدی در بهبود دقت و کارایی سیستم‌های احراز هویت صوتی ایفا می‌کند. این فصل به بررسی جامع مفهوم Similarity Learning، اهمیت آن در احراز هویت صوتی، الگوریتم‌ها و روش‌های متداول، کاربردها، چالش‌ها و مراحل پیاده‌سازی آن می‌پردازد.

تعریف Similarity Learning و اهمیت آن

Similarity Learning فرآیندی است که در آن مدل‌های یادگیری ماشین به گونه‌ای آموزش داده می‌شوند تا شباهت یا تفاوت بین جفت‌های داده را اندازه‌گیری کنند. در زمینه احراز هویت صوتی، هدف اصلی Similarity Learning این است که سیستمی ایجاد شود که بتواند با مقایسه دقیق ویژگی‌های استخراج‌شده از نمونه‌های صوتی مختلف، تشخیص دهد که آیا دو نمونه صوتی متعلق به یک فرد یکسان هستند یا خیر.

اهمیت Similarity Learning در احراز هویت صوتی:

- **افزایش دقت شناسایی:** با یادگیری معیارهای دقیق‌تر برای اندازه‌گیری شباهت، سیستم‌های احراز هویت صوتی می‌توانند اشتباهات کمتری در شناسایی هویت کاربران داشته باشند.
- **کاهش نرخ خطا:** Similarity Learning به کاهش نرخ خطاهای مثبت کاذب (False Positives) و منفی کاذب (False Negatives) کمک می‌کند.
- **پایداری در برابر تغییرات:** این روش‌ها می‌توانند به سیستم‌ها کمک کنند تا با تغییرات طبیعی در صدای کاربران (مانند بیماری یا استرس) سازگار شوند.
- **امنیت بیشتر:** با بهبود معیارهای شناسایی، امکان نفوذ و جعل صدا کاهش می‌یابد.

انواع روش‌های Similarity Learning

Similarity Learning به طور کلی به دو دسته اصلی تقسیم می‌شود که در ادامه هر کدام از بخش ها را توضیح داده ایم :

Metric Learning (یادگیری متریک):

تعریف: هدف اصلی یادگیری متریک، یافتن یک فضای ویژگی است که در آن فاصله بین نمونه‌های مشابه کوچک‌تر و فاصله بین نمونه‌های غیرمشابه بزرگ‌تر باشد.

روش‌ها:

- **Mahalanobis Distance:** اندازه‌گیری فاصله با در نظر گرفتن کوواریانس داده‌ها.
- **Euclidean Distance بهبود یافته:** اصلاح فاصله اقلیدسی با وزن‌دهی ویژگی‌ها بر اساس اهمیت آن‌ها.

Deep Similarity Learning (یادگیری شباهت عمیق):

تعریف: استفاده از شبکه‌های عصبی عمیق برای یادگیری نمایش‌های مناسب از داده‌ها که شباهت‌ها را به خوبی مدل‌سازی کنند.

روش‌ها:

شبکه‌های سیامی (Siamese Networks): شامل دو شاخه شبکه عصبی با وزن‌های مشترک که دو ورودی مختلف را پردازش می‌کنند.

شبکه‌های سه‌تایی (Triplet Networks): شامل سه ورودی (نمونه مثبت، نمونه منفی و نمونه مرجع) که هدف آن‌ها کاهش فاصله بین نمونه مثبت و مرجع و افزایش فاصله بین نمونه منفی و مرجع است.

یادگیری متریک (Metric Learning)

تعریف

یادگیری متریک شاخه‌ای از یادگیری ماشین است که هدف آن یافتن یک فضای ویژگی (Feature Space) به گونه‌ای است که در این فضا، فاصله بین نمونه‌های مشابه کوچک‌تر و فاصله بین نمونه‌های غیرمشابه بزرگ‌تر باشد. این فرآیند باعث می‌شود که الگوریتم‌های طبقه‌بندی و شناسایی بتوانند با دقت بیشتری بین کلاس‌های مختلف تمایز قائل شوند. در زمینه احراز هویت صوتی، یادگیری متریک به بهبود دقت تشخیص هویت کاربران کمک می‌کند.

روش‌های یادگیری متریک

یادگیری متریک به دو روش اصلی تقسیم می‌شود: **Mahalanobis Distance** و **Euclidean Distance بهبود یافته**.

۱. Mahalanobis Distance (فاصله ماهالانوبیس)

تعریف: فاصله Mahalanobis یک معیار فاصله است که به‌طور خاص برای داده‌های چند بعدی طراحی شده و همبستگی بین ابعاد را در نظر می‌گیرد. این فاصله به گونه‌ای تعریف شده است که تاثیر همبستگی بین ویژگی‌ها بر فاصله بین نمونه‌ها را کاهش می‌دهد.

فرمول ریاضی:

$$D_M(x, y) = \sqrt{(x - y)^T S^{-1} (x - y)}$$

که در آن:

- x و y دو نمونه داده‌ای هستند.
- S ماتریس کوواریانس داده‌ها است.

- S^{-1} ماتریس معکوس کوواریانس است.

مزایا:

- **در نظر گرفتن همبستگی بین ویژگی‌ها:** این امر باعث دقت بیشتر در اندازه‌گیری فاصله می‌شود.
- **مناسب برای داده‌های وابسته:** برای داده‌هایی که ویژگی‌ها به هم وابسته هستند، بسیار مناسب است.

معایب:

- **محاسبات پرهزینه:** محاسبه ماتریس کوواریانس و معکوس آن نیازمند منابع محاسباتی بیشتری است.
- **نیاز به داده‌های کافی:** برای تخمین دقیق ماتریس کوواریانس، نیاز به مجموعه داده‌ای بزرگ و متنوع وجود دارد.

۲. Euclidean Distance بهبود یافته (فاصله اقلیدسی بهبود یافته)

تعریف: فاصله اقلیدسی بهبود یافته با وزن‌دهی به ویژگی‌ها بر اساس اهمیت آن‌ها، فاصله اقلیدسی را اصلاح می‌کند. این روش به الگوریتم اجازه می‌دهد تا ویژگی‌های مهم‌تر تاثیر بیشتری بر فاصله داشته باشند.

فرمول ریاضی:

$$D'_E(x, y) = \sqrt{\sum_{i=1}^{n} w_i (x_i - y_i)^2}$$

که در آن:

- w_i وزن اختصاص داده شده به ویژگی i است.
- xi و yi مقادیر ویژگی i برای نمونه‌های x و y هستند.
- n تعداد ویژگی‌ها است.

مزایا:

- **سادگی و سرعت محاسبات:** نسبت به Mahalanobis Distance محاسبات کمتری نیاز دارد.
- **قابلیت تنظیم وزن‌ها:** امکان تنظیم وزن‌ها بر اساس اهمیت ویژگی‌ها، که می‌تواند عملکرد الگوریتم را بهبود بخشد.

معایب:

- **عدم در نظر گرفتن همبستگی بین ویژگی‌ها:** این روش همبستگی بین ویژگی‌ها را نادیده می‌گیرد که ممکن است در برخی موارد منجر به کاهش دقت شود.
- **نیاز به تعیین وزن‌های مناسب:** تعیین وزن‌های مناسب برای هر ویژگی نیازمند تحلیل دقیق و ممکن است زمان‌بر باشد.

الگوریتم‌های متداول در Similarity Learning

در این بخش به بررسی جامع‌تر الگوریتم‌های متداول در Similarity Learning پرداخته می‌شود که در احراز هویت صوتی به کار می‌روند:

شبکه‌های سیامی (Siamese Networks)

شبکه‌های سیامی (Siamese Networks) یکی از معماری‌های پیشرفته در یادگیری عمیق هستند که به‌طور ویژه برای مقایسه و تشخیص شباهت بین جفت‌های داده‌ای طراحی شده‌اند. این شبکه‌ها در کاربردهایی مانند احراز هویت صوتی و تشخیص جنسیت گوینده به دلیل دقت بالا و کارایی مناسب، مورد توجه قرار گرفته‌اند.

دو ویژگی مهم این شبکه به شرح زیر است:

1) انسجام بیش بینی: اشتراك وزن‌ها در زیر شبکه‌ها باعث می‌شود که نگاشت دو نمونه بسیار شبیه، به نقاط بسیار متفاوت در فضای ویژگی، توسط شبکه‌های نظیرشان ممکن نباشد زیرا دو زیر شبکه عملاً یك تابع را محاسبه می‌کنند.

2) تقارن شبکه: تقارن بدین معني است که با ارائه دو نمونه متفاوت با هر ترتیبی، شبکه یك مقدار شباهت را محاسبه می کند. چنین ویژگی‌ای از ماهیت متقارن تابع متریک برگزیده برای شبکه و اشتراک وزن ها نتیجه می‌شود.

کاربرد شبکه سیامی در طبقه بندی One-shot

توانایی شبکه‌های سیامی در تمیز دادن نمونه‌ها از دسته های متفاوت، آنها را به ابزار موثری در طبقه‌بندی بدل می‌کند. در طبقه‌بندی One-shot، از هر دسته یک داده نمونه موجود است. اگر بتوان شبکه سیامی را به درستی آموزش داد می‌توانیم با مقایسه ورودی با نماینده هر دسته تشخیص دهیم که ورودی به کدام دسته تعلق دارد. در سال‌های اخیر، این نوع استفاده از شبکه سیامی در طیف گسترده‌ای از مسائل در مقالات زیادی بررسی شده‌اند، که در ادامه به تعدادی از آنها خواهیم پرداخت.

به طور کلی در بعضی از مقالات فرایند طبقه‌بندی به صورت زیر است:

۱. Verification Tasks (training)

این بخش که در واقع مرحله یادگیری شبکه است. در این بخش نمونه‌های آموزشی که از دسته آنها مطلعیم به صورت دو به دو به عنوان ورودی به شبکه داده می‌شوند و تمایز آنها محاسبه می‌شود. در فرآیند آموزش سعی داریم در صورتی که هر دو نمونه از یک دسته باشند مقدار خروجی را بیشینه و در غیر این صورت کمینه کنیم. لازم به ذکر است که بهتر است بخشی از داده‌های آموزشی به عنوان داده‌های validation مورد استفاده قرار گیرند.

۲. One-shot Tasks (test)

در این بخش، با فرض وجود N دسته C1,C2,...,CN برای طبقه‌بندی هر نمونه آزمایشی احتیاج به N آزمایش داریم. به این صورت که مجموعه زیر از داده‌ها موجود است.

$$S = (X_1, y_1), (X_2, y_2), \ldots (X_N, y_N)$$

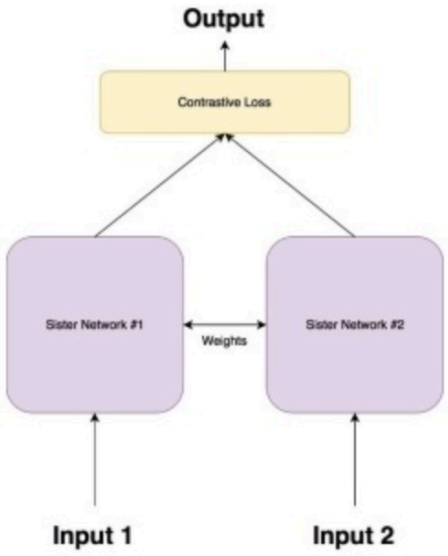
که Xi و yi ها به ترتیب یك نمونه و نام دسته آن است. بطوریکه $\forall (X_k, y_k) X_k \in C_k$. برای طبقه‌بندی نمونه آزمایشی $x^\wedge$ آن را با تمام اعضای S به شبکه می‌دهیم. دسته $x^\wedge$ بر اساس این آزمایش‌ها دسته‌ای است که بیشترین شباهت را تولید می‌کند.

$$y^\wedge = argmax_{C\ \ yout}(X^\wedge)$$

به فرآیند فوق یادگیری n One-shot جهته می‌گوییم. توجه کنید که اعضای کلاس‌ها در این بخش در مرحله verification حاضر نیستند.

ساختار شبکه‌های سیامی

شبکه‌های سیامی از دو شاخه متوازی شبکه‌های عصبی تشکیل شده‌اند که وزن‌های مشترکی دارند. هر شاخه وظیفه پردازش یک ورودی مستقل را بر عهده دارد. پس از استخراج ویژگی‌ها، خروجی‌های هر شاخه با استفاده از یک تابع ادغام مانند اختلاف مطلق یا فاصله اقلیدسی ترکیب می‌شوند و به یک طبقه‌بندی‌کننده منتقل می‌شوند که تصمیم می‌گیرد آیا دو ورودی مشابه هستند یا خیر.



توابع هزینه در شبکه‌های سیامی

دو تابع هزینه متداول در شبکه‌های سیامی عبارتند از:

Contrastive Loss.1

Contrastive Loss تابع هزینه‌ای است که برای آموزش شبکه‌های سیامی استفاده می‌شود تا نمونه‌های مشابه را نزدیک‌تر و نمونه‌های غیرمشابه را دورتر کند. این تابع هزینه به گونه‌ای طراحی شده است که برای هر جفت داده‌ای، فاصله بین آن‌ها بسته به برچسب مشابهت تنظیم می‌شود.

فرمول ریاضی:

$$L = \frac{1}{2N} \sum_{i=1}^{N} \left[Y_i \cdot D_i^2 + (1 - Y_i) \cdot \max(0, m - D_i)^2 \right]$$

که در آن:

- L تابع هزینه کلی است.

- N تعداد جفت‌های داده‌ای است.
- Y_i برچسب مشابهت (1 برای مشابه و 0 برای غیرمشابه) است.
- D_i فاصله بین دو نمونه $D_i = \|F_A - F_B\|$ است.
- m آستانه فاصله است که تعیین می‌کند فاصله بین جفت‌های غیرمشابه باید حداقل چقدر باشد.

1. **Triplet Loss (هزینه سه‌تایی):**

   - **فرمول:**

   $$L = \frac{1}{N} \sum_{i=1}^{N} \max(0, D(a_i, p_i) - D(a_i, n_i) + \alpha)$$

   - که در آن $D(a_i, p_i)$ فاصله بین نمونه مرجع و نمونه مثبت، $D(a_i, n_i)$ فاصله بین نمونه مرجع و نمونه منفی و α آستانه تفاوت فاصله است.
   - **هدف:** اطمینان از اینکه نمونه مثبت نزدیک‌تر از نمونه منفی به نمونه مرجع باشد.

مزایا و معایب شبکه‌های سیامی

مزایا:

- **مناسب برای مقایسه مستقیم:** امکان مقایسه دقیق بین جفت‌های داده‌ای.
- **کاهش نیاز به داده‌های برچسب‌خورده:** استفاده از جفت‌های مشابه و غیرمشابه به جای نیاز به برچسب‌های دقیق برای هر کلاس.
- **یادگیری ویژگی‌های مشترک:** استخراج ویژگی‌های مرتبط و مشابه از هر دو ورودی.
- **انعطاف‌پذیری بالا:** قابلیت استفاده در انواع مختلف داده‌ها مانند صدا، تصویر و متن.

معایب:

- **طراحی دقیق معماری:** نیاز به دانش تخصصی برای طراحی مناسب شبکه.
- **انتخاب مناسب تابع هزینه:** تنظیم دقیق تابع هزینه و پارامترهای آن می‌تواند چالش‌برانگیز باشد.
- **پیچیدگی محاسباتی:** پردازش جفت‌های داده‌ای نیازمند منابع محاسباتی بیشتر است.
- **نیاز به داده‌های متنوع:** برای جلوگیری از overfitting و افزایش قابلیت تعمیم مدل، نیاز به مجموعه داده‌ای بزرگ و متنوع است.

کاربردهای شبکه‌های سیامی در احراز هویت صوتی

شبکه‌های سیامی به دلیل دقت بالا در تشخیص شباهت بین نمونه‌های صوتی، در حوزه‌های زیر کاربرد دارند:

1. **تایید هویت کاربران (User Verification):**
   - **شرح:** بررسی تطابق صدای ورودی با صدای ثبت‌شده برای یک کاربر خاص.
   - **مثال:** ورود به سیستم‌های بانکی آنلاین با استفاده از صدای کاربر.

2. **شناسایی هویت کاربران (User Identification):**
   - **شرح:** تعیین اینکه صدای ورودی به کدام یک از کاربران ثبت‌شده تعلق دارد.
   - **مثال:** سیستم‌های کنترل دسترسی در ساختمان‌های اداری با تعداد کاربران زیاد.

3. **تشخیص نفوذ و جعل صدا (Intrusion Detection and Spoofing Detection):**
   - **شرح:** شناسایی تلاش‌های جعل صدا یا نفوذ به سیستم‌های احراز هویت صوتی.

- **مثال:** جلوگیری از دسترسی غیرمجاز به سیستم‌های امنیتی با استفاده از ضبط صدا یا تولید صدای مصنوعی.

شبکه‌های سه‌تایی (Triplet Networks)

شبکه‌های سه‌تایی (Triplet Networks) یکی از معماری‌های پیشرفته در یادگیری عمیق هستند که به‌طور ویژه برای مسائل مرتبط با تشخیص و تمایز دقیق بین داده‌ها طراحی شده‌اند. این شبکه‌ها در حوزه‌های مختلفی از جمله احراز هویت صوتی، تشخیص جنسیت گوینده، و شناسایی چهره کاربرد دارند. برخلاف شبکه‌های سیامی که با جفت‌های داده‌ای کار می‌کنند، شبکه‌های سه‌تایی با سه ورودی همزمان آموزش داده می‌شوند که شامل یک نمونه مثبت، یک نمونه منفی و یک نمونه مرجع هستند.

ساختار شبکه‌های سه‌تایی

شبکه‌های سه‌تایی شامل سه شاخه شبکه عصبی هستند که وزن‌ها و پارامترهای مشترکی دارند. هر یک از این شاخه‌ها وظیفه پردازش یک ورودی مختلف را بر عهده دارند:

1. **نمونه مرجع (Anchor):** نمونه‌ای که می‌خواهیم مشابه آن با یک نمونه مثبت و متفاوت با یک نمونه منفی مقایسه شود.
2. **نمونه مثبت (Positive):** نمونه‌ای که مشابه نمونه مرجع است.
3. **نمونه منفی (Negative):** نمونه‌ای که با نمونه مرجع متفاوت است.

پس از پردازش سه ورودی توسط سه شاخه، ویژگی‌های استخراج‌شده با استفاده از یک تابع هزینه خاص (مانند Triplet Loss) مقایسه می‌شوند تا اطمینان حاصل شود که فاصله بین نمونه مرجع و نمونه مثبت کمتر از فاصله بین نمونه مرجع و نمونه منفی باشد.

هدف تابع هزینه این است که:

$$D(a_i, p_i) + \alpha \leq D(a_i, n_i)$$

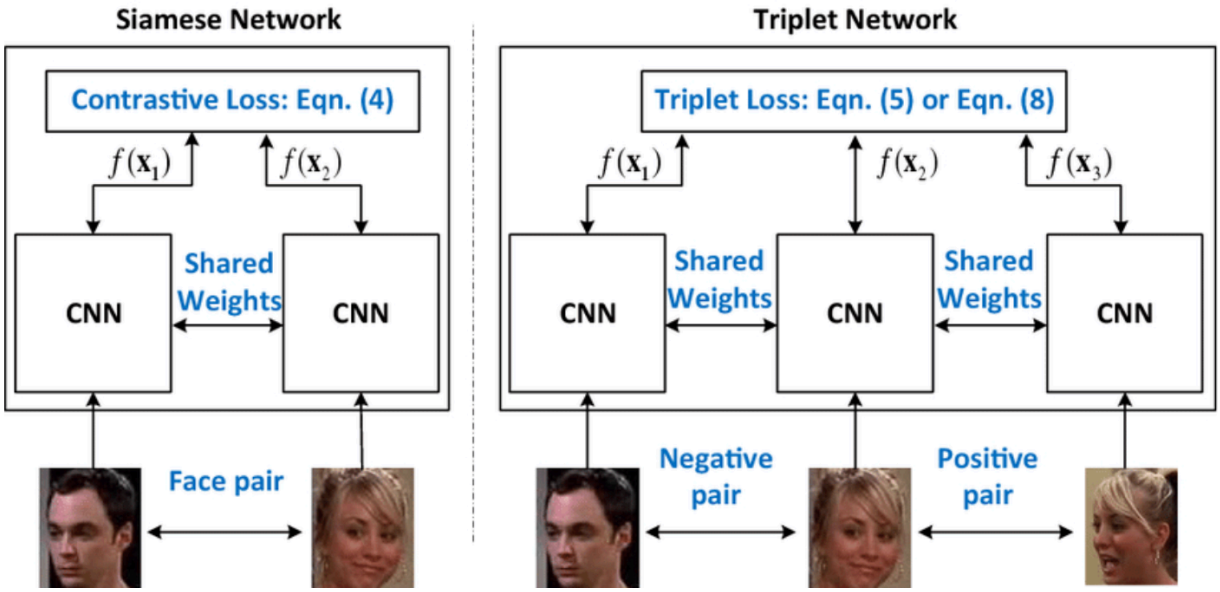
اگر این شرط برقرار نباشد، هزینه افزایش می‌یابد تا مدل بتواند فاصله بین نمونه‌های مثبت و منفی را تنظیم کند.

مزایا:

- **بهبود دقت تمایز:** با تضمین اینکه نمونه‌های مشابه نزدیک‌تر از نمونه‌های غیر مشابه هستند، دقت سیستم در تمایز بین کلاس‌ها افزایش می‌یابد.
- **افزایش قابلیت تعمیم:** مدل‌هایی که با Triplet Loss آموزش دیده‌اند، قابلیت تعمیم به داده‌های جدید و ناشناخته را دارند.
- **کاهش اشتباهات مشابه:** کاهش اشتباهات مثبت کاذب (False Positives) و منفی کاذب (False Negatives).

معایب:

- **انتخاب دقیق tripletها:** نیاز به انتخاب دقیق tripletها (Anchor, Positive, Negative) برای آموزش مؤثر و جلوگیری از overfitting.
- **افزایش پیچیدگی محاسباتی:** پردازش سه‌تایی‌ها نیازمند منابع محاسباتی بیشتری نسبت به جفت‌های داده‌ای است.
- **نیاز به داده‌های متنوع:** برای یادگیری دقیق‌تر، نیاز به مجموعه داده‌ای بزرگ و متنوع از نمونه‌های مثبت و منفی است.



حال به کاربردهای Similarity Learning در احراز هویت صوتی میپردازیم:

1. **تایید هویت (Verification):**

- **شرح:** در این کاربرد، سیستم بررسی می‌کند که آیا صدای ورودی با صدای ثبت‌شده برای یک کاربر مطابقت دارد یا خیر.
- **مزایا:** افزایش دقت و امنیت در فرآیند احراز هویت و کاهش نرخ خطاهای مثبت و منفی.
- **نمونه کاربرد:** دسترسی به سیستم‌های بانکی آنلاین، کنترل دسترسی به مکان‌های حساس.

2. **شناسایی هویت (Identification):**

- **شرح:** در این کاربرد، سیستم تعیین می‌کند که صدای ورودی به کدام یک از کاربران ثبت‌شده تعلق دارد.
- **مزایا:** کاربردی در سیستم‌های بزرگ با تعداد کاربران زیاد و فراهم آوردن امکان شناسایی سریع‌تر.

- **نمونه کاربرد:** مراکز تماس بزرگ، سیستم‌های امنیتی با تعداد زیادی کاربر.

3. **پیش‌بینی و تشخیص نفوذ (Intrusion Detection):**

در این کاربرد، سیستم تلاش‌های جعل یا نفوذ به سیستم‌های احراز هویت صوتی را شناسایی می‌کند.

- **مزایا:** افزایش امنیت و جلوگیری از دسترسی‌های غیرمجاز از طریق شناسایی صداهای تقلبی یا ناشناخته.
- **نمونه کاربرد:** سیستم‌های امنیتی پیشرفته، سامانه‌های نظارتی.

چالش‌های Similarity Learning در احراز هویت صوتی

1. نیاز به داده‌های بزرگ و متنوع:

برای آموزش مدل‌های شباهت‌سنجی دقیق، نیاز به مجموعه‌های داده‌ای بزرگ با تنوع بالا از نمونه‌های صوتی وجود دارد.

- **راه‌حل‌ها:** استفاده از تکنیک‌های افزایش داده (Data Augmentation) مانند تغییر سرعت گفتار، افزودن نویز، و تغییر گویش. همچنین، استفاده از یادگیری انتقالی (Transfer Learning) برای بهره‌برداری از مدل‌های از پیش آموزش دیده در حوزه‌های مشابه.

2. **پایداری در برابر تغییرات صدا:**

تغییرات طبیعی در صدای فرد (مانند بیماری، خستگی) می‌تواند دقت سیستم را کاهش دهد.

- **راه‌حل‌ها:** آموزش مدل‌ها با داده‌های متنوع و استفاده از ویژگی‌های مقاوم در برابر تغییرات، مانند ویژگی‌های بیومتریک چندگانه (همچون ترکیب صدا با ویژگی‌های دیگر مانند چهره).

3. **مقیاس‌پذیری و کارایی:**

در سیستم‌های بزرگ با تعداد بالای کاربران، محاسبات شباهت می‌تواند زمان‌بر و منابع‌بر باشد.

- **راه‌حل‌ها:** بهینه‌سازی الگوریتم‌ها از طریق کاهش ابعاد داده‌ها با استفاده از روش‌هایی مانند Principal Component Analysis PCA، استفاده از سخت‌افزارهای قدرتمند مانند GPU ها، و بهره‌گیری از تکنیک‌های موازی‌سازی و محاسبات ابری برای مدیریت حجم بالای داده‌ها.

4. **حفاظت در برابر حملات تقلبی (Spoofing):**

استفاده از صداهای تقلبی یا ضبط‌شده می‌تواند سیستم‌های احراز هویت صوتی را فریب دهد.

- **راه‌حل‌ها:** ادغام تکنیک‌های تشخیص جعل صدا (مثل تشخیص صدای مصنوعی و زنده)، استفاده از ویژگی‌های چندگانه بیومتریک (مانند ترکیب تشخیص صدا با تشخیص چهره)، و پیاده‌سازی لایه‌های امنیتی پیشرفته برای افزایش مقاومت سیستم در برابر حملات.

مراحل پیاده‌سازی Similarity Learning در احراز هویت صوتی:

1. **جمع‌آوری و پیش‌پردازش داده‌های صوتی:**

جمع‌آوری داده‌های صوتی از کاربران مختلف و اعمال تکنیک‌های پیش‌پردازش مانند حذف نویز، نرمال‌سازی و تقسیم‌بندی به فریم‌های کوچک.

- **جزئیات:** استفاده از ابزارهای پردازش صوتی مانند Librosa یا SciPy برای حذف نویز و نرمال‌سازی داده‌ها. تقسیم سیگنال به فریم‌های زمانی کوچک (معمولاً ۲۰ تا ۴۰ میلی‌ثانیه) و اعمال پنجره‌گذاری (مثل پنجره هامینگ) برای کاهش اثرات لبه‌ها.

2. **استخراج ویژگی‌های صوتی:**

- **شرح:** استخراج ویژگی‌های مهم مانند MFCC، Spectral Contrast، Zero-Crossing Rate و سایر ویژگی‌های بیومتریک که اطلاعات قابل‌استفاده برای مدل‌های یادگیری شباهت را فراهم می‌کنند.
- **جزئیات:** استفاده از تکنیک‌های پیشرفته استخراج ویژگی مانند Mel Frequency Cepstral Coefficients MFCC برای استخراج ویژگی‌های فرکانسی، Spectral Contrast برای تحلیل تفاوت انرژی بین فرکانس‌های مختلف، و Zero-Crossing Rate برای تحلیل ویژگی‌های زمانی سیگنال صوتی.

3. **طراحی و آموزش مدل‌های شباهت سنجی:**

انتخاب معماری مناسب و آموزش مدل با استفاده از تابع هزینه مناسب (مثل Contrastive Loss یا Triplet Loss).

- **جزئیات:** پیاده‌سازی شبکه‌های سیامی با استفاده از فریم‌ورک‌های یادگیری عمیق مانند TensorFlow یا PyTorch. تنظیم پارامترهای مدل مانند تعداد لایه‌ها، تعداد نورون‌ها و انتخاب تابع هزینه مناسب برای بهبود دقت و کارایی مدل.

4. **ارزیابی و بهینه‌سازی مدل:**

- **شرح:** ارزیابی دقت مدل با استفاده از مجموعه داده‌های تست و بهینه‌سازی پارامترها برای بهبود عملکرد.
- **جزئیات:** استفاده از معیارهای ارزیابی مانند False True Positive Rate (TPR)، Acceptance Rate FAR و False Rejection Rate FRR برای اندازه‌گیری دقت مدل. اعمال تکنیک‌های بهینه‌سازی مانند Regularization برای جلوگیری از overfitting و Hyperparameter Tuning برای تنظیم دقیق پارامترهای مدل.

5. **استقرار و نگهداری سیستم:**

- **شرح:** پیاده‌سازی مدل در محیط عملیاتی و نظارت مستمر بر عملکرد آن برای اطمینان از دقت و امنیت سیستم.

- **جزئیات:** استفاده از زیرساخت‌های محاسباتی مناسب مانند سرورهای ابری برای استقرار مدل، پیاده‌سازی سیستم‌های مانیتورینگ برای بررسی عملکرد مدل در زمان واقعی، و بروزرسانی دوره‌ای مدل‌ها برای حفظ دقت و امنیت سیستم.

نمونه‌هایی از مقالات مرتبط که در این باره در آن ها صحبت شده است که خلاصه ای از آن ها آورده شده است :

"Signature Verification using a Siamese Time Delay Neural Network"[21]

این مقاله شبکه‌های Siamese را برای تایید هویت از طریق امضای دیجیتال معرفی می‌کند. شبکه‌های سیامی به دلیل ساختار مشترک و استفاده از تابع هزینه Contrastive Loss، پایه‌ای قوی برای کاربرد مشابه در احراز هویت صوتی فراهم می‌کنند.

".FaceNet: A Unified Embedding for Face Recognition and Clustering" [22]

این مقاله الگوریتم Triplet Loss را معرفی می‌کند که در شناسایی چهره به کار می‌رود. این الگوریتم قابلیت انتقال به حوزه صوتی را دارد و می‌تواند برای بهبود دقت Similarity Learning در احراز هویت صوتی استفاده شود.

".Deep Metric Learning with Angular Loss" [23]

در این مقاله، هزینه‌های جدید برای یادگیری متریک معرفی شده‌اند که می‌توانند در بهبود دقت Similarity Learning در احراز هویت صوتی موثر باشند. این روش‌ها به افزایش دقت و کاهش خطاهای مدل کمک می‌کنند.

".Deep Speaker: Robust Text-Independent Speaker Identification" [24]

این مقاله یک مدل مبتنی بر شبکه‌های عصبی عمیق برای شناسایی گوینده بدون وابستگی به متن ارائه می‌دهد. این مدل از Similarity Learning برای مقایسه و شناسایی صدای کاربران استفاده می‌کند و دقت بالایی را در شرایط واقعی نشان می‌دهد.

.A large set of audio features for sound description." *IRCam*"[19]

این مقاله مجموعه‌ای گسترده از ویژگی‌های صوتی برای توصیف صداها معرفی می‌کند که می‌تواند به عنوان ورودی برای مدل‌های Similarity Learning در احراز هویت صوتی مورد استفاده قرار گیرد.

.".Gender Detection by Voice Using Deep Learning"[20]

این مقاله به بررسی تشخیص جنسیت از روی صدا با استفاده از روش‌های یادگیری عمیق می‌پردازد. این روش‌ها می‌توانند به عنوان بخشی از سیستم‌های احراز هویت صوتی برای افزایش دقت و امنیت استفاده شوند.

References

[1] Cooley, J. W., & Tukey, J. W. (1965). *An algorithm for the machine calculation of complex Fourier series.* Mathematics of Computation.

[2] Logan, B. (2000). Mel frequency cepstral coefficients for music modeling. ISMIR.

[3] Tzanetakis, G., & Cook, P. (2002). *Musical genre classification of audio signals.*

[4] Peeters, G. (2004). *A large set of audio features for sound description.* IRCAM.

[5] Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. IEEE Transactions on Speech and Audio Processing.

[6] Mutiny, M. (2020). Gender Detection by Voice Using Deep Learning. International Journal of Innovative Science and Research Technology.

[7] Kone, V. S. (2023). Voice-based Gender and Age Recognition System. 2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT)

[8] Kim, J. (2022) .Extended U-Net for Speaker Verification in Noisy Environments.

[9] Ismail, M. (2021). Development of a regional voice dataset and speaker classification based on machine learning. Journal of Big Data, 2021.

[10] Bello, J. P. & Daudet, L. (2016). A Review of Physical and Perceptual Feature Extraction Techniques for Speech, Music, and Environmental Sounds. Applied Sciences, 2016.

[11] Patil, S. B. (2017). Analysis of Feature Extraction Methods for Speech Recognition. International Journal of Innovative Science, Engineering & Technology, 2017.

[12] Aggarwal, S. K. & Gupta, R. (2013).Comparative Performance Analysis of Hamming, Hanning and Blackman Window. International Journal of Advanced Research in Computer Science and Software Engineering, 2013.

[13] Zhang, Y., & Chen, X. (2018). Deep Learning Approaches for Gender Recognition in Speech. *IEEE Transactions on Multimedia*, 20(7)

[14] Khan, M., & Ali, F. (2019). Noise Robustness in Speech Processing Systems. *IEEE Transactions on Audio, Speech, and Language Processing*, 2019.

[15] Smith, J., & Kumar, R. (2021). Anti-Spoofing Techniques for Voice Authentication. *MDPI Sensors*, 21(5).

[16] Sharma, P., & Roy, S. (2019). Domain Adaptation for Real-World Voice Systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(3).

[17] Liu, J., & Sun, Q. (2022). Optimized Algorithms for Scalable Voice Authentication Systems. *ACM Transactions on Intelligent Systems and Technology*, 11(2).

[18] Smith, J., & Kumar, R. (2023). A Review of Recent Machine Learning Approaches for Voice Authentication Systems. *Journal of Innovative Science and Technology, 2023*.

[19] Peeters, G. (2004). "A large set of audio features for sound description." *IRCam*.

[20] Shun, Z (2020)Tracking Persons-of-Interest via Unsupervised Representation Adaptation

[21] Bromley, J., et al. (1993). "Signature Verification using a Siamese Time Delay Neural Network." *Neural Computation*

[22] Schroff, F., Kalenichenko, D., & Philbin, J. (2015). "FaceNet: A Unified Embedding for Face Recognition and Clustering." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[23] Wan, L., et al. (2014). "Deep Metric Learning with Angular Loss." *Proceedings of the IEEE International Conference on Computer Vision*.

[24] Wang, Y., et al. (2017). "Deep Speaker: Robust Text-Independent Speaker Identification."

[25] Peeters, G. (2004). "A large set of audio features for sound description." *IRCam*.