به نام خدا

مبانی یادگیری ماشین

دکتر ابوالقاسمی - دکتر اعرابی

پروژه نهایی

سامان دوچی طوسی - 810101420

رضا چهرقانی - 810100401

علی دهقان‌زاده - 810101423

1.introduction to Voice Authentication

احراز هویت صوتی (Voice Authentication) یک روش امنیتی بیومتریک است که افراد را بر اساس ویژگی‌های منحصر به فرد صدایشان شناسایی می‌کند. این روش جنبه‌های مختلف گفتار فرد مانند لهجه، زیر و بمی صدا و سرعت را تحلیل می‌کند تا یک اثر صوتی منحصر به فرد ایجاد کند.

اهمیت احراز هویت صوتی در امنیت بالاتر و راحتی آن است:

1. نسبت به روش‌های سنتی مانند رمز عبور یا پین، امن‌تر است زیرا جعل یا هک آن دشوارتر است.
2. تجربه کاربری روان‌تری ارائه می‌دهد و اجازه احراز هویت از طریق گفتار طبیعی را می‌دهد.
3. دسترسی‌پذیری را برای افراد دارای معلولیت بهبود می‌بخشد.
4. کارایی عملیاتی را افزایش داده و هزینه‌های مرتبط با بازنشانی رمز عبور و پشتیبانی را کاهش می‌دهد.

کاربردهای احراز هویت صوتی شامل موارد زیر است:

1. احراز هویت موبایلی بدون دست، به ویژه در محیط‌هایی که سایر روش‌های بیومتریک نامناسب هستند، مانند خودروها.
2. احراز هویت برای دستگاه‌های تشخیص گفتار مانند Google Home یا Alexa آمازون.
3. تایید هویت در تماس‌های پشتیبانی مشتری، که نیاز به اشتراک‌گذاری اطلاعات شخصی را حذف می‌کند.
4. دسترسی امن به اطلاعات بانکی و انجام تراکنش‌ها.
5. افزایش امنیت در مراکز تماس، با برخی بانک‌ها که کاهش 50 درصدی در تقلب‌های تلفنی را گزارش کرده‌اند.
6. احراز هویت برای تماس‌های صوتی اینترنتی (VoIP)، خدمات دولت الکترونیک و خدمات اضطراری.

فناوری احراز هویت صوتی همچنان در حال تکامل است و پیشرفت‌ها در هوش مصنوعی و یادگیری ماشینی، دقت آن و توانایی تشخیص تلاش‌های متقلبانه را بهبود می‌بخشد.

[What is Authentication? | Definition from TechTarget]

در زمینه احراز هویت صوتی، مفاهیم close-set-authentication و open-set-authentication به دو رویکرد متفاوت در شناسایی و تایید هویت کاربران اشاره دارند:

Close-set-authentication

در این روش، سیستم از میان مجموعه محدود و مشخصی از کاربران شناخته شده، هویت فرد را تأیید می‌کند. به عبارت دیگر، تمام کاربران مجاز از قبل در سیستم ثبت شده‌اند.

ویژگی‌ها:

- محدود به تعداد مشخصی از کاربران
- دقت بالاتر در شناسایی
- مناسب برای سیستم‌های با تعداد کاربر محدود

Open-set-authentication

در این روش، سیستم باید قادر به تشخیص ورود کاربران ناشناس باشد. یعنی علاوه بر شناسایی کاربران مجاز، باید بتواند افراد خارج از مجموعه را نیز تشخیص دهد.

ویژگی‌ها:

- قابلیت مقابله با کاربران ناشناس
- پیچیدگی بیشتر در پیاده‌سازی
- مناسب برای سیستم‌های با تعداد کاربر نامحدود یا متغیر

تفاوت‌های اصلی

1. محدوده کاربران: Close-set محدود به کاربران شناخته شده است، در حالی که open-set قابلیت تشخیص کاربران ناشناس را دارد.
2. پیچیدگی: Open-set معمولاً پیچیده‌تر است زیرا نیاز به مکانیزم‌های اضافی برای تشخیص ورود کاربران ناشناس دارد.
3. کاربرد: Close-set برای سیستم‌های با تعداد کاربر ثابت و محدود مناسب است، در حالی که open-set برای سیستم‌های با تعداد کاربر متغیر یا نامحدود مناسب‌تر است.

پیاده‌سازی

:Close-set-authentication

1. جمع‌آوری نمونه‌های صوتی از تمام کاربران مجاز

2. استخراج ویژگی‌های صوتی و ایجاد مدل‌های صوتی برای هر کاربر
3. مقایسه صدای ورودی با مدل‌های موجود و انتخاب نزدیک‌ترین تطابق

:Open-set-authentication

1. انجام مراحل 1 و 2 مشابه close-set
2. ایجاد یک مدل پس‌زمینه یا مدل جهانی برای نمایش صداهای غیر مجاز
3. مقایسه صدای ورودی با مدل‌های کاربران و مدل پس‌زمینه
4. تصمیم‌گیری بر اساس آستانه‌های از پیش تعیین شده برای پذیرش یا رد هویت

کاربرد در Voice Authentication

در احراز هویت صوتی، انتخاب بین close-set و open-set به نوع کاربرد و نیازهای امنیتی سیستم بستگی دارد:

1. سیستم‌های داخلی سازمانی: معمولاً از close-set استفاده می‌کنند زیرا تعداد کاربران محدود و مشخص است.
2. سیستم‌های بانکی و مالی: اغلب از open-set استفاده می‌کنند تا بتوانند تلاش‌های غیرمجاز را تشخیص دهند.
3. سیستم‌های تلفنی مشتریان: معمولاً از open-set استفاده می‌کنند زیرا با تعداد زیادی از کاربران سر و کار دارند.
4. دستیارهای صوتی شخصی: می‌توانند از ترکیبی از هر دو روش استفاده کنند؛ close-set برای کاربران اصلی و open-set برای تشخیص صداهای ناآشنا.

انتخاب روش مناسب به فاکتورهایی مانند سطح امنیت مورد نیاز، تعداد کاربران، و منابع محاسباتی در دسترس بستگی دارد.

Difference Between Open and Closed Sets - Andrea Minini
OATH (Open Authentication) Definition | CardLogix Corporation
What is OAuth? How the open authorization framework works | CSO Online
Authentication methods | Login.gov

2. Challengesin Voice Authentication

در زمینه‌های احراز هویت صوتی (Voice Authentication) و طبقه‌بندی جنسیت (Gender Classification)، چالش‌های اساسی و راه‌حل‌های نوین به شرح زیر است:

چالش‌های اساسی

احراز هویت صوتی

1. حملات دیپ‌فیک: پیشرفت‌های سریع در هوش مصنوعی، به ویژه AI مولد، باعث ایجاد صداهای مصنوعی بسیار واقعی شده که تشخیص آن‌ها از صدای واقعی دشوار است.
2. مقیاس‌پذیری حملات: هوش مصنوعی امکان انجام حملات در مقیاس بزرگ را فراهم کرده است.
3. تغییرات صدا: عواملی مانند بیماری، استرس یا تغییرات فیزیولوژیکی می‌توانند صدای فرد را تغییر دهند.
4. نویز محیطی: صداهای پس‌زمینه می‌توانند دقت تشخیص را کاهش دهند.

طبقه‌بندی جنسیت

1. تنوع صدا: تفاوت‌های فرهنگی و زبانی در ویژگی‌های صوتی مرتبط با جنسیت.
2. افراد ترنس و غیر باینری: چالش در طبقه‌بندی صحیح افرادی که در طیف جنسیتی قرار دارند.

راه‌حل‌های نوین

1. استفاده از هوش مصنوعی و یادگیری عمیق: الگوریتم‌های پیشرفته برای افزایش دقت و مقاومت در برابر نویز.
2. تکنیک‌های حذف نویز: استفاده از روش‌های پردازش سیگنال برای بهبود کیفیت صدای ورودی.
3. احراز هویت مستقل از متن: تشخیص هویت بدون نیاز به عبارات از پیش تعیین شده.
4. ترکیب روش‌های بیومتریک: استفاده همزمان از چند روش مانند تشخیص چهره و صدا برای افزایش امنیت.
5. تشخیص حملات (Liveness Detection): استفاده از تکنیک‌های پیشرفته برای تشخیص صدای زنده از ضبط شده.
6. مدل‌های انطباق‌پذیر: سیستم‌هایی که می‌توانند با تغییرات تدریجی صدای کاربر سازگار شوند.

7. رویکردهای جامع در طبقه‌بندی جنسیت: استفاده از مدل‌های چندبعدی که طیف وسیع‌تری از هویت‌های جنسیتی را در نظر می‌گیرند.
8. آموزش با داده‌های متنوع: استفاده از مجموعه داده‌های گسترده و متنوع برای بهبود عملکرد در شرایط مختلف.

با توجه به چالش‌های موجود، برخی متخصصان پیشنهاد می‌کنند که استفاده از احراز هویت صوتی برای دسترسی به حساب‌های بانکی و اطلاعات حساس به تدریج کنار گذاشته شود.

2024 State of AI in the Speech Technology Industry: Voice Biometrics Both Profits From and Is Plagued by AI

OpenAI Recommends Phaseout of Voice-Based Authentication While Previewing Text-to-Speech Model | PYMNTS.com

Challenges in voice biometrics: Vulnerabilities in the age of deepfakes | ABA Banking Journal

Weighing the Benefits and Challenges of Voice Biometrics for Contact Centers: Part Two - Northridge Group

AI Voice Assistants: Scope, Benefits & Challenges in 2024

3. پیش پردازش داده های صوتی

a. بحث در مورد اهمیت پیش پردازش داده های صوتی در زمینه voice authentication و gender classification

پیش‌پردازش داده‌های صوتی نقش بسیار مهمی در زمینه‌های تشخیص هویت صوتی و طبقه‌بندی جنسیت ایفا می‌کند. این فرآیند شامل چندین مرحله کلیدی است که کیفیت و دقت تحلیل‌های بعدی را به طور قابل توجهی افزایش می‌دهد.

اهمیت پیش‌پردازش در تشخیص هویت صوتی:

- کاهش نویز: حذف صداهای ناخواسته و مزاحم از سیگنال صوتی، که باعث بهبود کیفیت صدا و افزایش دقت تشخیص می‌شود.
- استانداردسازی فرمت: تبدیل داده‌های صوتی به فرمت‌های استاندارد و یکسان، که پردازش و مقایسه آنها را آسان‌تر می‌کند.
- استخراج ویژگی: استخراج ویژگی‌های مهم صوتی مانند ضرایب کپسترال فرکانس مل (MFCCs) که برای شناسایی صدای افراد بسیار مفید هستند.

اهمیت پیش‌پردازش در طبقه‌بندی جنسیت:

- تحلیل فرکانس پایه: میانگین فرکانس پایه (f0) یکی از مهم‌ترین شاخص‌ها برای تشخیص جنسیت است. پیش‌پردازش مناسب امکان استخراج دقیق این ویژگی را فراهم می‌کند.
- نرمال‌سازی دامنه: این کار باعث می‌شود مدل‌های یادگیری ماشین نسبت به تفاوت‌های سطح صدا حساس نباشند و تمرکز بیشتری بر ویژگی‌های مرتبط با جنسیت داشته باشند.
- کاهش تغییرپذیری: حذف ویژگی‌های غیرمرتبط با جنسیت مانند لهجه یا احساسات، که می‌تواند دقت طبقه‌بندی را افزایش دهد.

به طور کلی، پیش‌پردازش مناسب داده‌های صوتی منجر به بهبود قابل توجه عملکرد سیستم‌های تشخیص هویت صوتی و طبقه‌بندی جنسیت می‌شود. این امر با افزایش کیفیت داده‌ها، کاهش نویز، و استخراج ویژگی‌های مرتبط امکان‌پذیر می‌شود.

Preprocessing the Audio Dataset - GeeksforGeeks
Audio signal processing and feature extraction | Deep Learning Systems Class Notes | Fiveable

b. توضیح مراحل پیش پردازش با تمرکز بر:

حذف نویز (Noise Reduction) یکی از مهم‌ترین مراحل در پیش‌پردازش داده‌های صوتی است. این فرآیند به بهبود کیفیت صدا و افزایش دقت در تحلیل‌های بعدی کمک می‌کند.

روش‌های سنتی حذف نویز:

- فیلترینگ طیفی (Spectral Gating): این روش معمول برای حذف نویز در موسیقی استفاده می‌شود. در این روش، سیگنال تنها در سطوح صدای بالا عبور داده می‌شود.
- فیلترهای صوتی: انواع مختلفی از فیلترها مانند فیلتر پایین‌گذر، بالاگذر، میان‌گذر و میان‌نگذر برای پالایش سیگنال صوتی قبل از پردازش‌های بیشتر استفاده می‌شوند.

روش‌های پیشرفته مبتنی بر یادگیری عمیق:

- شبکه‌های عصبی کانولوشنی (CNNs): معماری‌های مختلف CNN مانند VGG، ResNet و EfficientNet برای حذف نویز استفاده می‌شوند. این مدل‌ها قادر به استخراج ویژگی‌های پیچیده از داده‌های صوتی هستند.
- شبکه‌های مولد تخاصمی (GANs): این روش از یک مولد برای تولید صدای تمیز از ورودی‌های نویزی و یک تشخیص‌دهنده برای تمایز بین صدای واقعی و تولیدشده استفاده می‌کند.

تکنیک‌های پیشرفته پردازش سیگنال:

- نرمال‌سازی انرژی هر کانال (PCEN): این تکنیک انرژی سیگنال‌های صوتی را نرمال می‌کند و مقاومت مدل در برابر نویز را افزایش می‌دهد.
- تبدیل فوریه سریع (FFT): اندازه FFT نقش مهمی در اثربخشی تکنیک‌های حذف نویز دارد. FFT بزرگ‌تر برای نویزهای فرکانس پایین و طولانی‌مدت مناسب است، در حالی که FFT کوچک‌تر برای حذف صداهای گذرا مانند کلیک و پاپ بهتر عمل می‌کند.

Deep Learning Noise Reduction Techniques | Restackio
The Role of Adaptive Noise Reduction in Improving Audio Quality Using Adobe Audition | Noble Desktop

نرمال‌سازی فرآیندی است که دامنه سیگنال صوتی را تنظیم می‌کند تا به سطح مطلوبی برسد. دو نوع اصلی نرمال‌سازی عبارتند از:

- نرمال‌سازی قله (Peak Normalization): در این روش، بلندترین نقطه سیگنال به سطح مشخصی (معمولاً 0 دسی‌بل) تنظیم می‌شود.
- نرمال‌سازی بلندی (Loudness Normalization): این روش بر اساس درک شنوایی انسان، میانگین بلندی صدا را تنظیم می‌کند.

Audio normalization - Wikipedia

پنجره‌بندی (Windowing) یکی از مراحل مهم در پیش‌پردازش سیگنال‌های صوتی است. این تکنیک به تقسیم سیگنال به قطعات کوچک‌تر زمانی کمک می‌کند که برای تحلیل‌های دقیق‌تر ضروری است.

پنجره‌بندی معمولاً برای تحلیل طیفی استفاده می‌شود تا بتوان بخش کوتاهی از یک سیگنال طولانی‌تر را مشاهده و محتوای فرکانسی آن را تحلیل کرد. این روش به ویژه در پردازش سیگنال‌های گفتاری مهم است، زیرا صحبت از توالی فونم‌ها تشکیل شده و ویژگی‌های سیگنال در طول زمان تغییر می‌کند.

انواع توابع پنجره:

- پنجره مستطیلی (Rectangular): ساده‌ترین نوع پنجره که می‌تواند باعث ایجاد ناپیوستگی شود.
- پنجره هنینگ (Hanning): یک تابع سینوسی که به صفر می‌رسد و ناپیوستگی را حذف می‌کند.
- پنجره همینگ (Hamming): مشابه هنینگ، اما به صفر نمی‌رسد و ناپیوستگی جزئی ایجاد می‌کند.

انتخاب تابع پنجره مناسب:

- برای دقت در دامنه، پنجره Flat Top مناسب است.
- برای دقت در فرکانس، پنجره مستطیلی (در صورت عدم وجود ناپیوستگی) ایده‌آل است.
- پنجره‌های بلکمن، همینگ و هنینگ برای اکثر کاربردها عملکرد خوبی دارند.

طول پنجره معمولاً بین 20 تا 30 میلی‌ثانیه انتخاب می‌شود که برای تحلیل گفتار مناسب است. همپوشانی پنجره‌ها معمولاً با 30 تا 50 درصد همپوشانی اعمال می‌شوند تا از دست رفتن اطلاعات به حداقل برسد.

مزایای پنجره‌بندی شامل کاهش ناپیوستگی در ابتدا و انتهای سیگنال و کاهش نشت طیفی است. با این حال، باید توجه داشت که سیگنال نهایی ممکن است دقیقاً شبیه سیگنال اصلی نباشد

پنجره‌بندی مناسب به کاهش اثرات ناپیوستگی در لبه‌های سیگنال کمک می‌کند و دقت تحلیل‌های طیفی را افزایش می‌دهد. انتخاب نوع پنجره و پارامترهای آن باید با توجه به نوع سیگنال و هدف تحلیل انجام شود.

3.2. Windowing — Introduction to Speech Processing
Understanding FFTs and Windowing - NI
Windowing an audio signal | Max Cookbook

4. Feature extracting techniques

Mel-Frequency Cepstral Coefficients یا به اختصار MFCC نحوه ای برای نمایش طیف توان کوتاه‌مدت صدا هستند که ویژگی‌های مهم سیگنال‌های صوتی را به شکلی فشرده‌تر از نمایش‌های طیفی سنتی نشان می‌دهند. محاسبهٔ این ضرایب به این صورت است که:

1. در ابتدا یک فیلتر پیش‌تقویت به سیگنال صوتی اعمال می شود. این فیلتر فرکانس‌های بالاتر را amplify می کند تا با کاهش طبیعی انرژی فرکانس بالا در گفتار مقابله کند. همچنین این کار نسبت signal-to-noise را افزایش می دهد تا عملکرد را افزایش دهد.

2. سیگنال به فریم‌های همپوشان تقسیم می‌شود که معمولاً دارای دامنه زمانی حدود ۲۰ تا ۴۰ میلی‌ثانیه هستند. هر فریم یک بخش کوتاه از سیگنال صوتی را ضبط می‌کند.

3. برای کاهش نشت طیفی در تبدیل فوریه‌ای که بعداً انجام می‌شود، یک تابع پنجره (مانند هَمینگ یا هَنینگ) به هر فریم اعمال می‌شود. این باعث می شود از spectral leakage جلوگیری شود.

4. الگوریتم تبدیل فوریه سریع (FFT) به هر فریم پنجره‌شده اعمال می‌شود تا سیگنال دامنه زمانی به دامنه فرکانسی تبدیل شود. FFT عملاً الگوریتم مربوط به تبدیل فوریهٔ گسسته است.

5. محور فرکانس به مقیاس مِل تبدیل می‌شود، که واکنش گوش انسان به فرکانس‌های مختلف را تقلید می‌کند. این کار با استفاده از مجموعه‌ای از فیلترهای مثلثی که بر اساس مقیاس مِل فاصله دارند، انجام می‌شود که بر فرکانس‌های پایین‌تر تأکید و وضوح را در فرکانس‌های بالاتر کاهش می‌دهد. (سیستم شنیداری انسان اینگونه است که تغییرات در فرکانس های پایین را بسیار راحت تر از تغییرات در فرکانس بالا تشخیص می دهد. به عنوان مثال، تغییر فرکانس از 100 هرتز به 200 هرتز در مقایسه با تغییر فرکانس از 1500 هرتز به 1600 هرتز بسیار راحت تر تشخیص داده می شود. برای همین فرکانس را به مقیاس مل تبدیل می کنیم.) فرمول آن به صورت زیر است:

$$M = 2595\, log_{10}(1 + f/700)$$

6. توان خروجی هر فیلتر بانک معمولاً به مقیاس لگاریتمی تبدیل می‌شود. این مرحله ویژگی‌های ادراکی بلندی صدا را تقلید می‌کند.

7. در پایان، یک تبدیل کسینوسی گسسته به طیف لگاریتمی مِل اعمال می‌شود تا MFCCها تولید شوند. این مرحله ضرایب را غیر همبسته می‌کند و ابعاد را کاهش می‌دهد، و امکان ارائه فشرده‌ای از ویژگی‌های طیفی را فراهم می‌کند.

ضرایبی که بدست می آیند معمولاً به عنوان یک بردار از ضرایب ارائه می‌شوند که ضرایب اولیه اطلاعات مهم‌تری درباره صدا را به خود اختصاص می‌دهند، در حالی که ضرایب بالاتر جزئیات بیشتری را ارائه می‌دهند.به عنوان مثال ضریب صفرم معمولا انرژی میانگین سیگنال را در خود دارد.

Fast Fourier Transform یا به اختصار FFT الگوریتمی برای تبدیل فوریه گسسته است که در مقایسه با الگوریتم عادی تبدیل فوریه گسسته، سرعت بسیار بالاتری دارد. از این تبدیل برای یافتن فرکانس غالب در صدا، تغییر فرکانس خاص در صدا (مانند زیر یا بم در موسیقی)، تغییر تن و سرعت سیگنال صدا، کاهش نویز و ... استفاده می شود. دلیل سرعت بسیار بالای آن این است که صدا به بخش های کوچکتری تقسیم می شود که از ویژگی های symmetry و periodity استفاده شود. نتیجهٔ نهایی یک آرایه ای از اعداد مختلط است که برای هر فرکانس در سیگنال، اندازه و فاز آن را در خود دارد. سیگنال‌های صوتی معمولاً با استفاده از STFT (short-time fourier transform) پردازش می‌شوند که FFT را بر روی قاب‌های کوتاه و همپوشان داده‌های صوتی اعمال می‌کند. یک اندازه رایج قاب FFT ۱۰۲۴ نمونه است که تعادلی بین وضوح فرکانسی و موقعیت‌یابی زمانی ایجاد می‌کند. تحلیل FFT یک طیف از باکت‌های فرکانسی تولید می‌کند که هر کدام نشان‌دهنده دامنه یک مولفه فرکانسی خاص در سیگنال صوتی است. در کاربردهای آکوستیک، باکت‌های FFT اغلب به اکتاوها نرمال می‌شوند که منجر به یک خط صاف برای نویز صورتی می‌شود.

Log Mel Spectrogram نمایشی از یک سیگنال صوتی است که مفاهیم مقیاس مل و مقیاس لگاریتمی را ترکیب می‌کند. سیگنال صوتی ابتدا به فریم‌های همپوشان تقسیم می‌شود و به هر فریم یک تبدیل فوریه کوتاه‌مدت (STFT) اعمال می‌شود. این فرآیند، سیگنال دامنه زمانی را به دامنه فرکانسی تبدیل می‌کند و یک نمایش با مقادیر مختلط از محتوای فرکانسی سیگنال در طول زمان ارائه می‌دهد.

از STFT، طیف دامنه برای هر فریم محاسبه می‌شود که دامنه مولفه‌های مختلف فرکانسی موجود در سیگنال صوتی را نشان می‌دهد. سپس طیف دامنه از یک فیلتر بانک مل عبور داده می‌شود که شامل فیلترهای مثلثی همپوشان است که بر اساس مقیاس مل فاصله دارند. این گام نمایشی فرکانسی را به یک نمایشی در مقیاس مل تبدیل می‌کند و فرکانس‌هایی را که برای ادراک انسانی بیشتر مرتبط هستند، تاکید می‌کند.

پس از اعمال فیلتر بانک مل، خروجی معمولاً به مقیاس لگاریتمی تبدیل می‌شود. این تبدیل لگاریتمی نحوه ادراک انسان از بلندی صدا را تقلید می‌کند، به‌طوری‌که نسبت‌های برابر از شدت به‌عنوان تفاوت‌های برابر در بلندی ادراک می‌شوند. لگاریتم به فشرده‌سازی دامنه پویا سیگنال صوتی کمک می‌کند و تحلیل و پردازش آن را آسان‌تر می‌کند. نتیجه یک نمایش دو بعدی است که در آن یک محور زمان (فریم‌ها) و محور دیگر باندهای فرکانسی مل را نشان می‌دهد. هر مقدار در این نمایش به انرژی لگاریتمی سیگنال صوتی در یک باند فرکانسی مل خاص در یک زمان خاص مربوط می‌شود.

از خود spectral نیز ویژگی های بسیار زیادی می توان استخراج کرد:

Spectral Centroid: فرکانسی است که در آن بیشتر انرژی سیگنال متمرکز شده است. اگر طیف سیگنال را به‌صورت یک نمودار تصور کنید، مرکز ثقل طیفی (spectral centroid) نقطه‌ای است که در آن "تعادل انرژی" فرکانس‌ها برقرار است و از فرمول زیر محاسبه می شود.

$$Spectral\ centroid\ = \frac{\sum_i f_i |X(f_i)|}{\sum_i X(f_i)}$$

که اگر عدد آن بزرگ باشد یعنی در فرکانس های بالا باشد، صدا زیر و شفاف تر و اگر در فرکانس های پایین تر باشد، صدا بم و عمیق تر است.

Spectral contrast: این ویژگی بر روی فرکانس های اکسترمم تمرکز دارد، مخصوصا بر روی تفاوت انرژی بیشترین و کمترین فرکانس در یک بازه. برای محاسبه کنتراست طیفی، ابتدا سیگنال صوتی به باندهای فرکانسی تقسیم می‌شود. سپس انرژی هر باند محاسبه می‌شود و کنتراست بین باندهای فرکانسی مختلف به دست می‌آید. معمولاً از یک فیلتر بانک برای استخراج باندهای فرکانسی استفاده می‌شود و سپس کنتراست به عنوان تفاوت بین انرژی باندهای فرکانسی بالا و پایین محاسبه می‌شود. تفسیری که می توان داشت به این صورت است که کنتراست بالا نشان‌دهنده تفاوت زیاد بین انرژی فرکانس‌های بالا و پایین است. معمولاً در صداهایی با ساختار هارمونیک قوی (مانند موسیقی یا صداهای پیچیده) دیده می‌شود. همچنین کنتراست طیفی پایین نشان‌دهنده یکنواختی انرژی در طیف است. معمولاً در نویز سفید یا صداهای ساده رخ می‌دهد.

Spectral Roll-Off: رول‌آف طیفی معیاری است که فرکانسی را نشان می‌دهد که در زیر آن درصد معینی (معمولاً ۸۵٪ یا ۹۵٪) از کل انرژی طیفی قرار دارد. این ویژگی برای تمایز بین محتوای هارمونیک و غیر هارمونیک در یک سیگنال استفاده می‌شود. اینگونه که ابتدا طیف دامنه سیگنال محاسبه می‌شود. توزیع انرژی تجمعی محاسبه می‌شود. در نهایت فرکانس باندی شناسایی می‌شود که در آن انرژی تجمعی به درصد مشخصی از کل انرژی می‌رسد.

فرکانس رول‌آف بالا نشان‌دهنده وجود محتوای فرکانس بالا است که معمولاً به صداهای ضربه‌ای یا نویزی اشاره دارد. در مقابل فرکانس رول‌آف پایین نشان‌دهنده تسلط محتوای فرکانس پایین است که معمولاً در صداهای هارمونیک یا تونال مشاهده می‌شود.

استخراج ویژگی تشخیص گفتار از غیرگفتار - ایران متلب
Mel-frequency Cepstral Coefficients (MFCC) for Speech Recognition - GeeksforGeeks
Spectral FeaturesUnderstanding the Mel Spectrogram | by Leland Roberts | Analytics Vidhya | Medium

Chroma Features: شدت هر یک از ۱۲ کلاس نت C، C#، D، D#، E، F، F#، G، G#، A، A#، B را در یک بخش صوتی مشخص می‌کنند. این یعنی ویژگی‌های کرومایی (Chroma Features)بر محتوای هارمونیک سیگنال صوتی تمرکز دارند و نه محتوای فرکانسی خاص آن.

سیگنال صوتی ابتدا به دامنه فرکانسی با استفاده از تبدیل فوریه کوتاه‌مدت تبدیل می‌شود. این شامل تقسیم سیگنال صوتی به فریم‌های همپوشان و اعمال تبدیل فوریه به هر فریم است. سپس ستون های فرکانسی به‌دست‌آمده از STFT به ۱۲ کلاس نت نقشه‌برداری می‌شوند. این کار با جمع کردن انرژی باین‌های فرکانسی که به هر کلاس نت مربوط می‌شوند، انجام می‌شود. به عنوان مثال، تمام فرکانس‌های مربوط به C، C#، D و غیره جمع می‌شوند تا یک مقدار برای هر کلاس نت ایجاد شود.

پس از آن وکتور کرومایی chroma vector به‌دست‌آمده می‌تواند نرمال‌سازی شود تا تغییرات در بلندی صدا را در نظر بگیرد و اطمینان حاصل کند که نمایندگی ویژگی در نمونه‌های مختلف صوتی یکسان است.

ویژگی‌های کرومایی معمولاً به‌صورت یک وکتور ۱۲ بعدی نمایش داده می‌شوند، که هر بعد به یکی از ۱۲ کلاس نت مربوط می‌شود. مقادیر در این وکتور نمایانگر انرژی یا شدت هر کلاس نت در فریم زمانی تحلیل‌شده هستند.

از آنجا که ویژگی‌های کرومایی بر اساس کلاس‌های نت و نه فرکانس‌های خاص هستند، در برابر تغییرات در گام (مانند انتقال) مقاوم هستند. همچنین ویژگی‌های کرومایی تحلیل از نظر محاسباتی را بسیار کارآمد تر می سازند.

Chroma feature - Wikipedia
Zero-crossing rate - Wikipedia
Linear Predictive Coding
(PDF) Feature extraction methods LPC, PLP and MFCC in speech recognition

نرخ عبور از صفر (Zero-Crossing Rate) به فرکانسی اشاره دارد که سیگنال صوتی ورودی، که تابعی از زمان است، محور صفر را عبور می‌کند. این عبور زمانی اتفاق می‌افتد که شکل موج از مثبت به منفی یا برعکس تغییر می‌کند. برای محاسبه نرخ عبور از صفر، دامنه سیگنال صوتی نظارت می‌شود و تعداد دفعاتی که سیگنال از صفر عبور می‌کند، شمارش می‌شود. معمولاً سیگنال صوتی به فریم‌هایی با مدت زمان مشخص تقسیم می‌شود و ZCR برای هر فریم محاسبه می‌شود. سپس نرخ با تقسیم تعداد عبور از صفر بر طول فریم محاسبه می‌شود.

یک ZCR بالا معمولاً نشان‌دهنده مقدار زیادی محتوای فرکانس بالا یا "نویز" در سیگنال است، در حالی که ZCR پایین نشان‌دهنده سیگنالی با محتوای فرکانس بالا کمتر است، مانند گفتار زنان یا سازهای موسیقی با صداهای خشن کمتر.

کدگذاری پیش‌بینی خطی (Linear Predictive coding) با بهره‌گیری از همبستگی‌های موجود در سیگنال، مقدار داده‌های مورد نیاز برای نمایش سیگنال گفتار را کاهش می‌دهد.

LPC سیگنال گفتار را با فرض اینکه گفتار توسط یک سیستم خطی زمان‌ناپایدار (LTI) تولید می‌شود، مدل‌سازی می‌کند. این سیستم مجرای صوتی است که توسط یک منبع نویز سفید یا پالس گلوتال تحریک می‌شود. برای

ایجاد مدل، LPC ضرایب بازتاب را از یک سری نمونه‌های گفتاری محاسبه می‌کند. این ضرایب ویژگی‌های شکل‌دهی مجرای صوتی را برای آن نمونه‌ها توصیف می‌کنند.

این ضرایب بازتاب یا ضرایب پیش‌بینی خطی (LPCs) برای نمایش فشار هوای در لب‌های مدولاتور استفاده می‌شوند، به جای ارسال سیگنال صوتی خام. این امر منجر به فشرده‌سازی داده‌ها می‌شود.

مدل پیش‌بینی خطی بر اساس پیش‌بینی مقدار فعلی سیگنال بر اساس ترکیب خطی از مقادیر گذشته است. این پیش‌بینی به عنوان معادله پیش‌بینی خطی شناخته می‌شود. ضرایب موجود در معادلات پیش‌بینی خطی به عنوان ضرایب بازتاب شناخته می‌شوند. این ضرایب قدرت پیش‌بینی را نمایان می‌سازند و برای تشکیل فیلتر استفاده می‌شوند.

نحوهٔ محاسبه اینگونه است که سیگنال گفتار به بلوک‌های کوچک و همپوشان تقسیم می‌شود. نمونه‌های هر بلوک بر اساس نمونه‌های گذشته پیش‌بینی می‌شوند و یک خطا (باقیمانده پیش‌بینی) محاسبه می‌شود. یک فیلتر با استفاده از ضریب بازتاب طراحی می‌شود تا باقیمانده پیش‌بینی با شکل طیفی سیگنال گفتار مطابقت داشته باشد. ضرایب بازتاب و گاهی اوقات باقیمانده‌های پیش‌بینی کوانتیزه و کدگذاری می‌شوند.

سیگنال‌های گفتار می‌توانند به شدت فشرده شوند، که آن‌ها را برای انتقال در کانال‌های با پهنای باند محدود مناسب می‌سازد. نمایش‌های LPC معمولاً در برابر نویز پس‌زمینه و اعوجاج‌های کانالی مقاوم هستند.

پیش‌بینی خطی ادراکی (Perceptual Linear Prediction) بر اساس اصول کدگذاری پیش‌بینی خطی (LPC) ساخته شده و جنبه‌های ادراکی شنوایی انسان را در نظر می‌گیرد. PLP شامل چندین مرحله است که رویکرد سنتی LPC را برای در نظر گرفتن عوامل ادراکی تغییر می‌دهد.

مشابه PLP ،LPC با یک مرحله پیش‌تأکید آغاز می‌شود تا اجزای فرکانس بالا در سیگنال گفتار تقویت شوند، که به بهبود نسبت سیگنال به نویز کمک می‌کند. سیگنال صوتی به فریم‌های همپوشان تقسیم می‌شود و STFT برای تبدیل سیگنال از دامنه زمان به دامنه فرکانس اعمال می‌شود. محور فرکانس به مقیاس مل تغییر شکل می‌دهد که حساسیت گوش انسان به فرکانس‌های مختلف را منعکس می‌کند. این مرحله فرکانس‌های پایین‌تر را تقویت کرده و وضوح در فرکانس‌های بالاتر را کاهش می‌دهد. مقیاس مل بیشتر پردازش می‌شود تا باندهای بحرانی شنوایی انسان را شبیه‌سازی کند. این شامل اعمال یک بانک فیلتر است که پاسخ فرکانسی گوش انسان را تقلید می‌کند.

خروجی بانک فیلتر با استفاده از مقیاس لگاریتمی تبدیل می‌شود تا نحوه ادراک بلندی صدا توسط انسان را شبیه‌سازی کند. این مرحله دامنه دینامیکی سیگنال را فشرده می‌کند. در نهایت، پیش‌بینی خطی بر روی طیف لگاریتمی اعمال می‌شود تا ضرایب PLP استخراج شوند. این ضرایب نمایانگر پوشش طیفی سیگنال گفتار هستند و ویژگی‌های ادراکی را در نظر می‌گیرند.

با گنجاندن عوامل ادراکی، PLP نمایی دقیق‌تر از سیگنال‌های گفتاری ارائه می‌دهد که با شنوایی انسان هم‌راستا است و منجر به بهبود عملکرد در وظایف شناسایی گفتار می‌شود. ویژگی‌های PLP معمولاً در برابر نویز و تغییرات در سبک گفتار نسبت به ویژگی‌های سنتی LPC مقاوم‌تر هستند

5. Similarity Learning

a. تحیلی Similarity Learning در حوزه تحلیل صدا و نحوه استفاده از آن برای تشخیص شباهت بین ویژگی های صوتی

Similarity Learning در حوزه تحلیل صدا روشی است که برای یادگیری و مقایسه شباهت‌های بین سیگنال‌های صوتی استفاده می‌شود. این تکنیک به طور گسترده در کاربردهایی مانند بازیابی ویدیوهای مشابه بر اساس صدا و جستجوی موسیقی مشابه مورد استفاده قرار می‌گیرد.

یکی از رویکردهای مهم در این زمینه، استفاده از یادگیری خودنظارتی (self-supervised learning) است. در این روش، داده‌های آموزشی با اعمال تغییرات تصادفی روی کلیپ‌های صوتی ایجاد می‌شوند.هدف این است که مدل یاد بگیرد نسبت به این تغییرات تقریباً ناوردا (invariant) باشد.

برای تشخیص شباهت بین ویژگی‌های صوتی، معمولاً از یک تابع رمزگذاری (encoding function) استفاده می‌شود که سیگنال‌های صوتی را به بردارهایی در یک فضای ویژگی تبدیل می‌کند. سپس فاصله بین این بردارها (مثلاً فاصله اقلیدسی) به عنوان معیاری برای شباهت استفاده می‌شود.

در یک نمونه کاربردی، شرکت Epidemic Sound از یک موتور تشخیص شباهت موسیقی استفاده می‌کند که شامل دو بخش است: یک مدل یادگیری ماشین که کلیپ‌های صوتی را به بردار تبدیل می‌کند و یک موتور جستجو.

برای بهبود عملکرد این سیستم‌ها، استراتژی‌های نمونه‌برداری مختلفی مورد بررسی قرار گرفته‌اند. نحوه انتخاب نمونه‌ها برای دسته‌های کوچک (mini-batches) و تابع خطا (loss function) می‌تواند تأثیر قابل توجهی بر عملکرد مدل و همگرایی آن در طول آموزش داشته باشد.

در مجموع، Similarity Learning در تحلیل صدا ابزاری قدرتمند برای درک و مقایسه ویژگی‌های پیچیده صوتی است که کاربردهای گسترده‌ای در صنعت موسیقی و پردازش صدا دارد.

MUSICAL AUDIO SIMILARITY WITH SELF-SUPERVISED CONVOLUTIONAL NEURAL NETWORKS
A sample of sampling strategies for audio similarity learning

b. معرفی و توضیح Loss functions رایج در Similarity Learning (مانند: Contrastive Loss و Triplet Loss و ...)

در حوزه Similarity Learning، چندین تابع خطا (Loss function) رایج وجود دارد که برای آموزش مدل‌ها به منظور یادگیری شباهت‌ها و تفاوت‌ها بین نمونه‌ها استفاده می‌شوند. دو مورد از مهم‌ترین این توابع عبارتند از:

Contrastive Loss: این تابع یکی از اساسی‌ترین توابع خطا در یادگیری تقابلی است. هدف اصلی این تابع، افزایش شباهت بین جفت‌های مثبت (نمونه‌هایی از یک کلاس) و کاهش شباهت بین جفت‌های منفی (نمونه‌هایی از کلاس‌های متفاوت) در فضای ویژگی آموخته شده است.

این تابع خطا معمولاً به صورت یک تابع خطای مبتنی بر حاشیه (margin-based loss) تعریف می‌شود، که در آن شباهت بین نمونه‌ها با استفاده از یک معیار فاصله مانند فاصله اقلیدسی یا شباهت کسینوسی اندازه‌گیری می‌شود. Contrastive Loss با جریمه کردن نمونه‌های مثبت که خیلی از هم دور هستند و نمونه‌های منفی که خیلی به هم نزدیک هستند، محاسبه می‌شود.

What is Contrastive Learning? A guide.
Full Guide to Contrastive Learning | Encord

Triplet Loss یکی دیگر از توابع خطای محبوب در یادگیری شباهت است که اولین بار در سال 2015 در مقاله FaceNet معرفی شد. این تابع خطا بر اساس مقایسه سه نمونه کار می‌کند:

1. Anchor: یک نمونه مرجع
2. Positive: نمونه‌ای با همان برچسب Anchor
3. Negative: نمونه‌ای با برچسب متفاوت از Anchor

هدف Triplet Loss این است که فاصله بین Anchor و Positive را کمینه کند، در حالی که فاصله بین Anchor و Negative را بیشتر از یک حد آستانه (margin) نگه دارد. به طور ریاضی، این تابع خطا به صورت زیر محاسبه می‌شود:

$$L = max(d(a,p) - d(a,n) + m, 0)$$

که در آن d تابع فاصله و a نمونه Anchor و p نمونه Positive و n نمونه Negative و m مقدار margin است.

Triplet Loss به مدل کمک می‌کند تا بازنمایی‌هایی را یاد بگیرد که نمونه‌های مشابه را نزدیک به هم و نمونه‌های متفاوت را دور از هم قرار می‌دهد، که این امر برای وظایفی مانند تشخیص چهره و بازیابی تصویر بسیار مفید است.

Triplet loss - Wikipedia
Triplet Loss - Advanced Intro - Qdrant
Triplet Loss in Tensorflow Similarity | by Ridadogrul | AI Mind