دانشکده مهندسی برق و کامپیوتر

داده افزایی برای هیجانات (احساسات) در گفتار

استاد راهنما:سید جلال ذهبی

متین فاضل

بهمن ۱۴۰۲

فهرست مطالب

فصل ۱

مقدمه

۱.۱ داده افزایی برای هیجانات (احساسات) در گفتار

تشخیص عواطف داده های صوتی SER به تشخیص خودکار احساسات و عواطف انسانی اشاره دارد. به عنوان زمینه تحقیقاتی مهم، تشخیص عواطف دادهای صوتی به سرعت در حال رشد است، همچنین دارای پتانسیل بهبود تعامل انسان و رایانه مبتنی بر صدا، مانند سیستم داخل خودرو برای درک وضعیت عاطفی رانندگان به هنگام ایجاد تغییرات ناگهانی است.

۲.۱ شرح مسئله

کمبود داده از عمده چالش های پیچیده در تشخیص هیجانات در گفتار است که در این سه مورد شرح داده شده است:

- مشکل اول فقدان دیتاست اصوات صوتی طبیعی است. دیتاست های کمی برای تحقیقات در این زمینه به اشتراک گذاشته شده است. مخصوصا، بسیاری از دیتاست های صوتی که در شرایط واقعی تولید شده اند به علت یکسری محدودیت های قانونی در دسترس عمومی قرار نگرفته اند.
- مشکل دیگر برچسب گذاری داده های صوتی است. از آنجایی که احساسات ابراز شده متفاوت هستند، دسته بندی آنها بسیار مهم است. با این حال به دلیل عدم قطعیت بالا دسته بندی و تحلیل آنها کاری به شدت زمان بر است.
- در نهایت، داده های صوتی در اکثر پایگاه های داده به صورت نا متعادلی بر روی احساسات توزیع میشوند. به طور کلی، تعداد جملات با احساسات خنثی بیشترین تعداد را در بدنه گفتاری دارد. با این حال، برای ارزیابی دقت طبقه بندی، یک پایگاه داده متعادل برای تجزیه و تحلیل نیاز است. علاوه بر این، اگر یک جمله با احساسات مختلف ضبط شود، قضاوت انسان در مورد احساس درک شده میتواند صرفا بر اساس محتوای عاطفی جمله بدون تاثیر محتوای واژگانی آن باشد.

۳.۱ مشارکت ها

شبکه های مولد متخاصم (GANs)، یک روش کارآمد برای تولید دیتا هستند. با استفاده از یک بازی متخاصم بین یک تشخیص دهنده و یک مولد، شبکه های مولد متخاصم برای تولید نمونه هایی که از داده های واقعی قابل تشخیص نیستند، آموزش میبینند. علاوه بر این دارای مشخصه های زیر هستند:

- شبکه های مولد متخاصم میتوانند توزیع های احتمالی با ابعاد بالا را در مسائل پیچیده دنیای واقعی بیاموزند.
- این شبکه ها را میتوان با داده ها از دست رفته آموزش داد، که برای یادگیری نیمه نظارتی مناسب است.
- شبکه های مولد متخاصم دارای خروجی های چندوجهی هستند، به این معنی که میتوانند چندین پاسخ صحیح مختلف تولید کنند و تنوع نمونه های تولید شده را افزایش دهند.

هدف این پروژه ارزیابی عملکرد SER زمانی است که دادههای آموزشی واقعی در فضای ویژگی با دادههای مصنوعی تولید شده توسط شبکه های مولد متخاصم افزایش میابد. به طور خاص، نیاز به طراحی یک مدل مبتنی بر GAN است که بردارهای ویژگی مصنوعی از گفته های احساسی مختلف را تولید کند، به طوری که عملکرد یک شبکه عصبی دستهبند با دریافت برداهای ویژگی واقعی و مصنوعی به عنوان دادهای آموزشی بهبود داده شود.

فصل ۲

مفاهیم و اصطلاحات علمی

۱.۲ شبکه های مولد تخاصمی

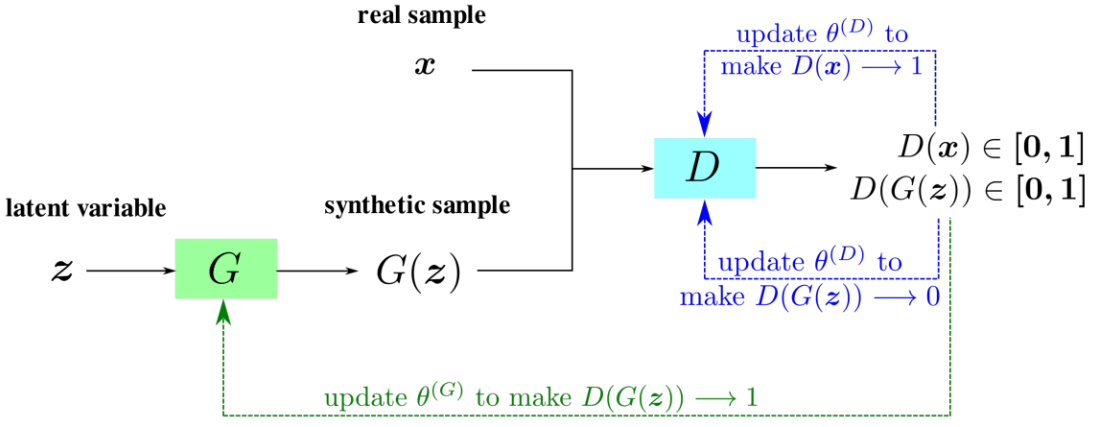
مدل های مولد به هر مدلی اطلاق می‌شود که مجموعه ای از نمونه‌های آموزشی گرفته‌شده از یک توزیع را یاد می‌گیرد که تخمینی از آن توزیع را نشان دهد. این شبکه‌ها بر اساس تولید داده به دو دسته صریح [۱] و ضمنی [۲] تقسیم می‌شوند. مدل های صریح تابع چگالی توزیع را مستقیماً محاسبه میکنند در حالی که مدل های ضمنی بر تولید نمونه‌هایی از توزیع ارائه شده توسط مدل تمرکز میکند.

۱.۱.۲ معماری

ک شبکه مولد متخاصمی، یک مدل مولد بر بر اساس نظریه بازی است که دو عامل را روبه‌روی هم قرار میدهد:

- یک تشخیص دهنده D [۳]
- یک مولد G [۴]

در شکل ۱.۱ طرح کلی یک شبکه مولد متخاصمی رسم شده است، که با نام Vanila GAN شناخته می‌شود. مولد به عنوان ورودی متغیر پنهان z را دریافت میکند که از یک مجموعه توزیع نویز $p_z(z)$ تهیه میشود و داده ساخته شده $G(z;\theta^G)$ را به عنوان خروجی تولید می کند. هدف مولد این است که یکسری داده تولید کند که از داده‌های واقعی غیر قابل تشخیص باشند. تشخیص دهنده از یک طرف داده های واقعی و از طرفی دیگر نمونه های مصنوعی $G(z)$ را به عنوان ورودی دریافت میکند. خروجی این مدل $D(x)$ یا $D(G(z))$ که احتمال واقعی بودن نمونه دریافتی را نشان می‌دهد. مولد تلاش می کند تا مقدار $D(G(z))$ را به عدد یک نزدیک کند تا تشخیص کننده را متقاعد کند که نمونه تولید شده شبیه به نمونه واقعی است، در حالی که تشخیص دهنده در تلاش است که مقدار $D(G(z))$ را صفر نزدیک کند و مقدار $D(x)$ را به یک نزدیک کند.



شکل ۱.۲: ساختار یک شبکه مولد متخاصمی

[1] Implicit
[2] Explicit
[3] Discriminator
[4] Generator

۲.۱.۲ تابع هزینه

در نظر داریم که به عنوان ورودی شبکه داده‌های $(s_i, y_i)_{i=1}^N$ را دریافت میکنیم که نیمی از آن داده های واقعی x و نیمی دیگر از آن داده‌های تولید شده $G(z)$ است. هر نمونه آموزشی s_i متناظر با یک برچسب y_i است. همه داده‌های واقعی دارای برچسب یک و تمامی داد های واقعی حاوی برچسب صفر می‌باشند. با توجه به اینکه هدف تشخیص دهنده یک دسته‌بند دودویی است، تابع هزینه آن به صورت یک binary cross-entropy تعیین می‌شود.

$$J^{(D)}(D,G) = H((s_i,y_i)_{i=1}^N, D) = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i \log(D(s_i)) + (1-y_i)\log(1-D(s_i))\right) \qquad (۱.۲)$$

اگر y_i را برابر با یک برای $s_i = x$ و برابر با صفر برای وقتی که $s_i = G(z)$ در نظر بگیریم، همچنین با جایگذاری میانگین ها به رابطه تابع هزینه تشخیص دهنده به عبارت زیر میرسیم:

$$J^{(D)}(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[1-\log D(G(z))] \qquad (۲.۲)$$

که p_{data} یک توزیع داده بر روی نمونه داده‌های واقعی x است. در یک بازی minimax که به نام بازی zero-sum نیز شناخته می‌شود، مجموع هزینه‌های تمامی بازیکنان همواره صفر است. که نشان می‌دهد تابع هزینه مولد مخالف $J^{(D)}$ است. در حالی که برای محاسبه تابع هزینه مولد در نزول گرادیانی عبارت دوم در معادله تابع هزینه اهمیت دارد. بنابراین تابع هزینه مولد در یک بازی minimax به صورت زیر تعریف شده است:

$$J^{(G)}(G) = \mathbb{E}_{z\sim p_z(z)}[1-\log D(G(z))] \qquad (۳.۲)$$

در حالت کلی این بازی در یک تابع ارزش خلاصه می‌شود که به صورت زیر تعریف می‌شود:

$$\min_G \max_D \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[1-\log D(G(z))] \qquad (۴.۲)$$

که مولد سعی در کمینه کردن آن و تشخیص دهنده سعی در بیشینه کردن آن دارد. در صورتی که تابع هزینه مولد در یک بازی minimax در واقعیت به خوبی عمل نمی‌کند، به دلیل اینکه وقتی مولد سعی در بیشینه کردن مقداری که تشخیص‌دهنده سعی در کمینه کردن آن را دارد، باعث می‌شود که تشخیص دهنده به آسانی تمامی داده‌های تولید شده توسط مولد را مصنوعی تشخیص دهد و آنها را رد کند. در نتیجه، نزول گرادیانی در مولد دچار اخلال می‌شود. مولد به جای آنکه مقدار تشخیص درست تشخیص دهنده را کمینه کند، مولد در یک بازی non-saturating heuristic سعی در بیشینه کردن اشتباه تشخیص دهنده دارد. و تابع هزینه آن در این بازی به صورت زیر تعریف می‌شود:

$$-\mathbb{E}_{x\sim p_{data}(x)}[\log D(G(x))] \qquad (۵.۲)$$

در شکل ۱.۲ اختلاف مقدار تابع هزینه در حالت minimax و non-saturating heuristic قابل مشاهده است. محور افقی احتمال قبول شدن یک نمونه داده مصنوعی به عنوان یک داده واقعی را نشان می‌دهد. هر چه مقدار این عدد بیشتر باشد، مولد میزان هزینه کمتری را میگیرد. قسمت سمت چپ تابع که میزان $D(G(z))$ نزدیک به صفر است، در ابتدای مرحله آموزش مدل اتفاق می افتد. در این زمان، تشخیص دهنده به راحتی می‌تواند تشخیص بدهد که داده نمونه متعلق به کدامین کلاس است.، زیرا مولد در ابتدا شروع به ساخت داده‌های مصنوعی با توجه به توزیع تصادفی $p_z(z)$ با یکسری پارامتر رندوم می‌کند. واضح است که منحنی بازی minimax .یک خط مستقیم است، که نشان می‌دهد مولد دارای گرادیان بسیار کمی است. با استفاده از نزول گرادیانی، مولد روند بهبود مدل را در مراحل ابتدایی متوقف کرده است. در نقطه مقابل، منحنی بازی non-saturating heuristic مقدار گرادیان خودش را در قسمت راست از دست می‌دهد، و در این نقطه بهینه نمونه داده‌های تولید شده قادر به گمراه کردن تشخیص دهنده هستند. بنابراین، بازی non-saturating معمولا در دنیای واقعی کاربرد بیشتری دارد و نسبتا بازی minimax جنبه نظری دارد.

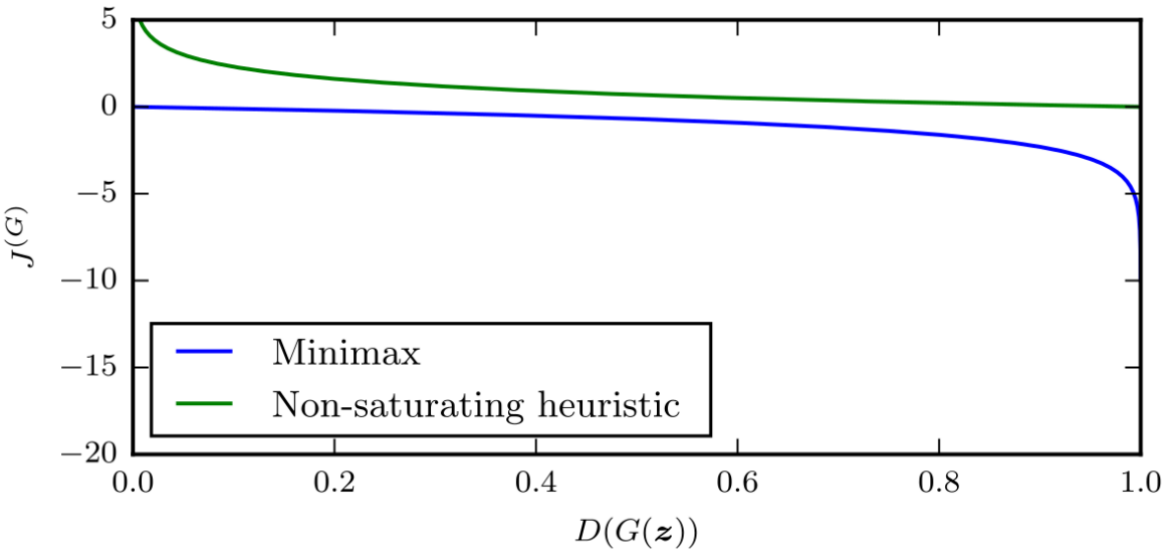
۳.۱.۲ انواع GAN برای تشخیص هیجانات در گفتار

در سال‌های گذشته شبکه های مولد تخاصمی برای تشخیص هیجانات در گفتار استفاده شده است. برای مثال در یک آزمایش از یک DCGAN برای تحلیل گفتار احساسی به روش نیمه نظارت شده استفاده کردند. در این قسمت ما تمرکز بر استفاده از شبکه های مولد متخاصمی برای تولید داده داریم، که در نهایت به ما کمک می‌کند که داده‌هایی تولید کنیم که شبیه به توزیع داده‌های واقعی در مسئله داده شده باشد. در این قسمت ما به تحلیل و بررسی سه نوع رایج شبکه های مولد متخاصمی می‌پردازیم که شامل: adversarial autoencoder, conditional GAN, and CycleGAN است.

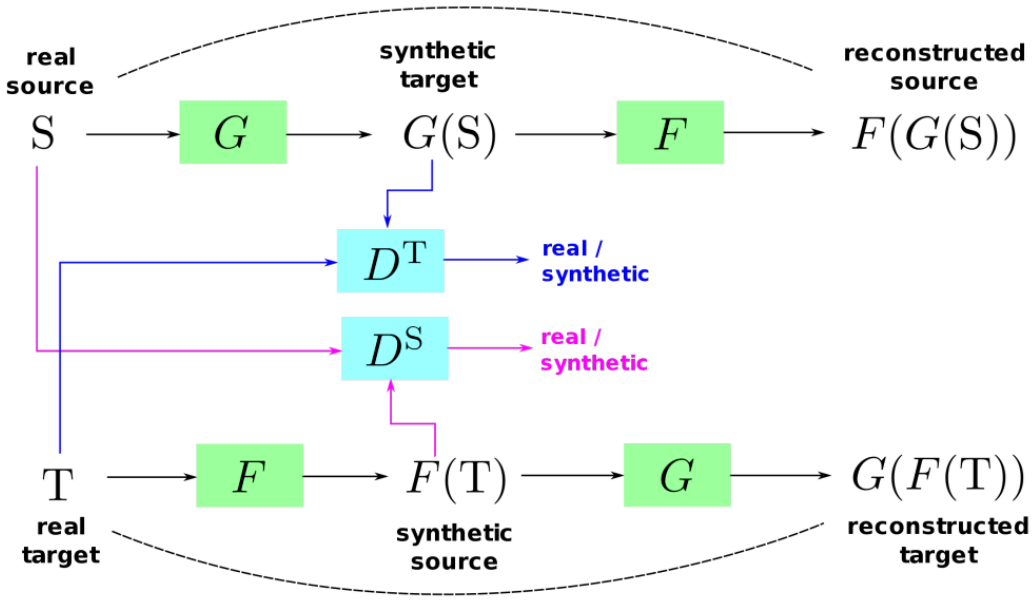
۴.۱.۲ شبکه های تخاصمی حلقوی پایدار

[1]شبکه های تخاصمی حلقوی پایدار که معروف به CycleGAN است یک روش بسیار موفق برای حل مسائل ترجمه متن به متن یا به کار گیری مجموعه داده‌های جفت نشده است. نگاشت دو طرفه ای که توسط شبکه CycleGAN آموخته می‌شود، می‌تواند ویژگی‌های خاص یک مجموعه تصویر را ثبت کند و بفهمد که چگونه می‌توان این ویژگی‌ها را به مجموعه تصویر دیگر ترجمه کند. موفقیت چشم‌گیر این نوع شبکه در انگیزه‌ای برای استفاده

[1]Cycle-consistent adversarial networks



شکل ۲.۲: مقایسه مقدار تابع هزینه مولد در بازی های minimax و non-saturating heuristic



شکل ۳.۲: ساختار CycleGan

در انتقال احساسات است. شکل ۱.۳ ساختار کلی یک CycleGAN را تصویر کشیده است. این شبکه حاوی دو تابع نگاشت G و F است. تابع G یادمی‌گیرد که داده‌های نمونه را از منبع S به دامنه هدف T تبدیل کند. همچنین، تابع نگاشت F نیز یک نگاشت معکوس نسبت به G است. هر دوی این توابع نگاشت G و F را می‌توان یکسری مولد برای هدف و منبع تولید داده در نظر گرفت. به علاوه، این شبکه دارای دو تشخیص دهنده D^T و D^S است. تشخیص دهنده D^T در مقابل مولد ،G وظیفه تشخیص واقعی بودن داده‌های تولید شده توسط G را برای داده‌های T دارد. همچنین، D^S تشخیص هم وظیفه تشخیص داده واقعی S را از مجموعه داده تولید شده F(T) دارد. همچنین در ادامه برای اطمینان حاصل کردن از اینکه تصویر ساخته شده قابلیت بازگردانی به داده نمونه اصلی را دارد، این شبکه سعی در ساخت نمونه داده‌های هدف و منبع را دارد به طوری که F(G(S)) باید شبیه به S و G(F(T)) باید شبیه به T باشد، که مربوط به قسمت پایداری این شبکه تخاصمی است.

میزان خطای CycleGAN تشکیل شده از یک خطای تخاصمی و خطای حلقوی پایدار است. خطای تخاصمی را می‌توانیم برای دو قسمت S و T در نظر بگیریم. خطای تخاصمی به صورت زیر محاسبه می‌شود:

$$\mathcal{L}_{GAN}(G, D^T, S, T) = \mathbb{E}_{t\sim p_t(t)}[1 - \log D^T(t)] + \mathbb{E}_{s\sim p_s(s)}[1 - \log 1 - D^T(G(s))] \quad (۶.۲)$$

$$\mathcal{L}_{GAN}(F, D^S, T, S) = \mathbb{E}_{s\sim p_s(s)}[1 - \log D^S(s)] + \mathbb{E}_{t\sim p_t(t)}[1 - \log 1 - D^S(G(t))] \quad (۷.۲)$$

قابل ذکر است که میزان خطای تخاصمی در قالب یک تابع ارزش بیان شده است. بنابراین حدف ما از این خطا $\min_G \max_{D^T} \mathcal{L}_{GAN}(G, D^T, S, T)$ و $\min_F \max_{D^S} \mathcal{L}_{GAN}(F, D^S, T, S)$ است.

با داشتن ظرفیت به اندازه بزرگ، شبکه می‌تواند همگی تصاویر ورودی را به هر ترتیب تصادفی از تصاویر در دامنه مقصد نگاشت کند. به گونه‌ای که هر یک از نگاشت‌های یادگرفته‌شده می‌تواند یک توزیع خروجی را ایجاد کند که با توزیع مقصد همخوانی داشته باشد. بنابراین این شبکه به یک تابع خطای دیگر هم نیاز دارد، که در پایین ذکر شده است:

$$\mathcal{L}_{cyc}(G, F) = \mathbb{E}_{t\sim p_t(t)}[||G(F(t)) - t||_1] + \mathbb{E}_{s\sim p_s(s)}[1 - \log 1 - D^T(G(s))] \quad (۸.۲)$$

در نتیجه به صورت کلی تابع هزینه به این صورت نمایش داده میشود:

$$\mathcal{L}(G, F, D^T, D^S) = \mathcal{L}_{GAN}(G, D^T, S, T) + \mathcal{L}_{GAN}(F, D^S, T, S) + \lambda\mathcal{L}_{cyc}(G, F) \quad (۹.۲)$$

۵.۱.۲ تقسیم‌بندی با رویکرد یادگیری ماشین یا یادگیری عمیق

یادگیری ماشینی برای تقسیم خودکار قسمت‌های مختلف یک تصویر استفاده می‌شود. معماری‌های U-Net در حل مسائل تقسیم‌بندی کارآمد هستند. در ادامه با توجه به اینکه روش استفاده شده مبتی بر یادگیری عمیق است ابتدا اصطلاحات و تئوری مربوط به آن را بررسی می‌نماییم.

۲.۲ سیگنال

سیگنال یک تغییر در یک کمیت معین در طول زمان است. برای صوت، مقداری که تغییر می کند فشار هوا است. ما می توانیم از فشار هوا در طول زمان نمونه برداری کنیم. سرعت نمونه برداری از داده ها می تواند متفاوت باشد، اما معمولاً ۱.۴۴ کیلوهرتز یا ۴۴۱۰۰ نمونه در ثانیه است. آنچه ما گرفته ایم یک شکل موج برای سیگنال است و می توان آن را با نرم افزار کامپیوتری تفسیر، اصلاح و تحلیل کرد.

۳.۲ سری فوریه

در ریاضیات، تبدیل فوریه[1] یک تبدیل ریاضیاتی است که توابعی را که بر حسب زمان یا فضا هستند، به توابعی بر حسب فرکانس زمانی یا فضایی تجزیه می‌کند، مانند بیان یک آکورد موسیقی بر حسب حجم‌ها و فرکانس‌های نت‌های تشکیل دهنده آن. اصطلاح تبدیل فوریه هم به نمایش دامنه فرکانس و هم به عملیات ریاضی مربوط به آن که نمایش دامنه فرکانس را به تابعی از مکان یا زمان مرتبط می‌کند گفته می‌شود.

تبدیل فوریه یک تابع از زمان، یک تابع مقدار مختلط از فرکانس است، که اندازه آن (قدر مطلق)، فرکانس موجود در تابع اصلی را نشان می‌دهد، و آرگومان آن اختلاف فاز سینوسی پایه در آن فرکانس است. تبدیل فوریه فقط محدود به توابع زمان نیست، اما به دامنه عملکرد اصلی، معمولاً دامنه زمان گفته می‌شود. معکوس تبدیل فوریه نیز وجود دارد که به صورت ریاضی تابع اصلی را از نمایش دامنه فرکانسی آن تولید می‌کند، که توسط قضیه عکس فوریه اثبات شده است.

۱.۳.۲ تعریف

تبدیل فوریه، نامیده شده به اسم ریاضیدان فرانسوی ژوزف فوریه، یک تبدیل انتگرالی است که هر تابع $f(t)$ را به یک تابع دیگر $F(\omega)$منعکس می‌کند. در این صورت، به $F(\omega)$ تبدیل فوریه تابع $f(t)$می‌گویند. حالت خاص تبدیل فوریه، سری فوریه نام دارد و آن زمانی کاربرد دارد که تابع $f(t)$ متناوب باشد، یعنی: $f(t+T) = f(t)$.چنانچه تابع متناوب نباشد یا به عبارتی، تناوب آن برابر بی‌نهایت باشد ($T \to \infty$)،از سری فوریه عبارت زیر به دست می‌آید:

$$F(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(t)e^{-i\omega t}dt \quad (۱۰.۲)$$

$$f(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} F(\omega)e^{i\omega t}d\omega \quad (۱۱.۲)$$

[1] Fourier transform

۴.۲ فاصله اولیه فریشه

فاصله اولیه فریشه[1] (FID) معیاری است که برای ارزیابی کیفیت تصاویر ایجاد شده توسط یک مدل مولد، مانند یک شبکه متخاصم مولد استفاده می‌شود. بر خلاف معیار قدیمی تر امتیاز اولیه (IS)، که فقط توزیع تصاویر تولید شده را ارزیابی می کند، فاصله اولیه فریشه توزیع تصاویر تولید شده را با توزیع مجموعه ای از تصاویر واقعی مقایسه می کند.

۱.۴.۲ تعریف

برای هر دو توزیع احتمالی u, v بر روی $\mathbb{R}^n$ که دارای میانگین و انحراف از معیار متناهی هستند، فاصله اولیه فریشه به صورت زیر تعریف می‌شود:

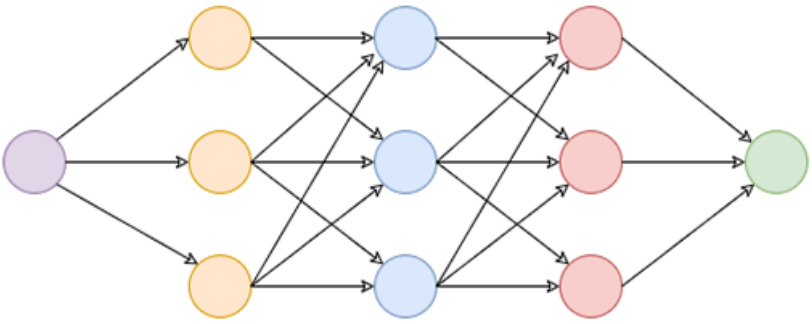
$$d_F(u,v) := \sqrt{inf_{\gamma \in \Gamma(u,v)} \int_{\mathbb{R}^n \times \mathbb{R}^n} ||x-y||^2 d\gamma(x,y)}, \tag{۱۲.۲}$$

به صورتی که $\Gamma(u,v)$ شامل تمامی مقادیر بر روی $\mathbb{R}^n \times \mathbb{R}^n$ است، با مقادیر حاشیه‌ای u, v به ترتیب بر روی عوامل اول و دوم. برای دو توزیع گاوسی چند بعدی، $N(\mu_1, \Sigma_1) N(\mu_1, \Sigma_1)$ این رابطه به صورت زیر است:

$$d_F(N(\mu_1,\Sigma_1), N(\mu_1,\Sigma_1))^2 = \|\mu_1 - \mu_2\|^2 + Tr\left(\Sigma_1 + \Sigma_2 - 2\sqrt{\Sigma_1 \Sigma_2}\right) \tag{۱۳.۲}$$

۵.۲ شبکه عصبی

شبکه‌های عصبی [2] از مغز ما الگو گرفته‌اند. نودهایی [3] وجود دارند که لایه‌ها [4] را در شبکه تشکیل می‌دهند و دقیقاً مانند نورون‌های مغز ما، نواحی مختلف را به هم متصل می‌کنند. به ورودی‌های نودها در یک لایه، وزنی اختصاص می‌یابد که تأثیری را که پارامتر بر نتیجه پیش‌بینی کلی دارد، تغییر می‌دهد. از آنجا که وزن‌ها به پیوندهای بین نودها اختصاص داده می‌شوند، ممکن است هر نود تحت‌تأثیر وزن‌های مختلف قرار گیرد. شبکه عصبی تمام داده‌های آموزش را در لایه ورودی می‌گیرد. سپس داده‌ها را از میان لایه‌های پنهان عبور داده، مقادیر را براساس وزن هر نود تغییر می‌دهد و در نهایت مقداری را در لایه خروجی برمی‌گرداند.



شکل ۴.۲: شبکه عصبی با چندین لایه پنهان. هر لایه چندین گره دارد.

تنظیم درست یک شبکه عصبی برای رسیدن به نتایج سازگار و قابل‌اعتماد ممکن است کمی زمان‌بر باشد. آزمایش و آموزش شبکه عصبی، یک فرآیند متعادل‌سازی برای تعیین مهم‌ترین ویژگی‌های مدل است.

۶.۲ شبکه عصبی پیچشی

شبکه عصبی پیچشی [5] نوع خاصی از شبکه عصبی با چندین لایه است که داده‌هایی را که آرایش شبکه‌ای دارند، پردازش کرده و سپس ویژگی‌های مهم آن‌ها را استخراج می‌کند. یک مزیت بزرگ استفاده از CNN ها این است که نیازی به انجام پیش‌پردازش زیادی روی تصاویر نیست.

در بیشتر الگوریتم‌هایی که پردازش تصویر را انجام می‌دهند، فیلترها معمولاً توسط یک مهندس بر اساس روش‌های اکتشافی (heuristic) ایجاد می‌شوند. CNN ها می‌توانند مهم‌ترین ویژگی فیلترها را بیاموزند و چون به پارامترهای زیادی احتیاج نیست، صرفه‌جویی زیادی در وقت و عملیات آزمون و خطا صورت می‌گیرد.

[1] Fréchet inception distance
[2] Neural Networks - NNs
[3] Nodes
[4] Layers
[5] Convolutional Neural Network - CNN

هدف اصلی الگوریتم CNN این است که با حفظ ویژگی‌هایی که برای فهم آنچه داده‌ها نشان می‌دهند مهم هستند، داده‌ها را به فرم‌هایی که پردازش آن‌ها آسان‌تر است، درآورد. آن‌ها همچنین گزینه خوبی برای کار با مجموعه داده‌های عظیم هستند.

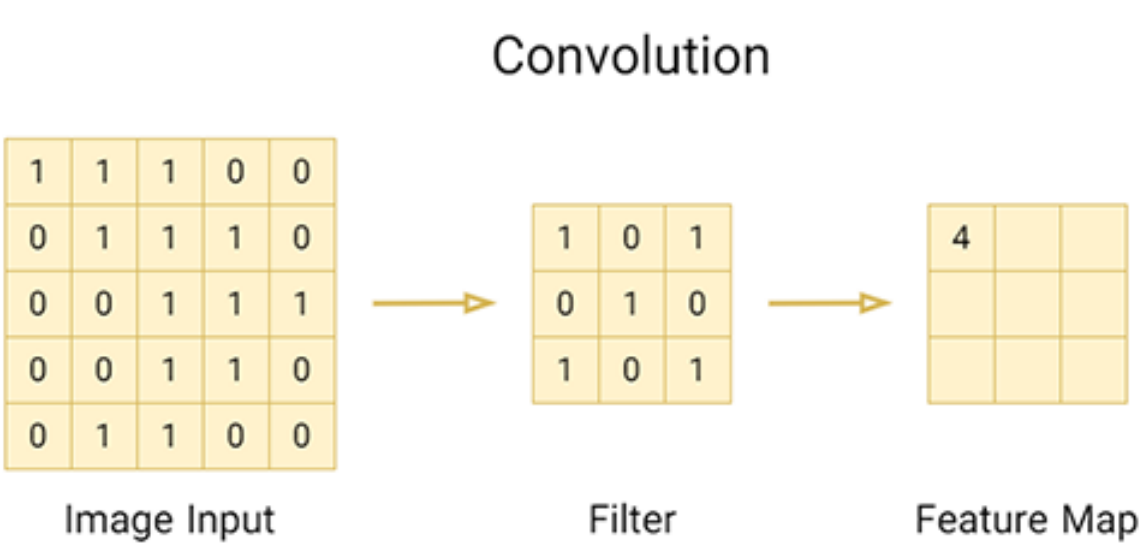
یک تفاوت بزرگ بین CNN و شبکه عصبی معمولی این است که CNN ها برای مدیریت ریاضیات پشت صحنه، از کانولوشن استفاده می‌کنند. حداقل در یک لایه از CNN ، به‌جای ضرب ماتریس از کانولوشن استفاده می‌شود. کانولوشن‌ها تا دو تابع را می‌گیرند و یک تابع را برمی‌گردانند.

CNN ها با اعمال فیلتر روی داده‌های ورودی شما کار می‌کنند. چیزی که آن‌ها را بسیار خاص می‌کند، این است که CNN ها می‌توانند فیلترها را هم‌زمان با فرایند آموزش، تنظیم کنند. به این ترتیب، حتی وقتی مجموعه داده‌های عظیمی مانند تصاویر داشته باشید، نتایج به‌خوبی و درلحظه دقیق‌تر می‌شوند.

از آنجا که می‌توان فیلترها را برای آموزش بهتر CNN تازه‌سازی کرد، نیاز به فیلترهای دستی از بین می‌رود و این انعطاف‌پذیری بیشتری در تعداد و ارتباط فیلترهایی که بر روی مجموعه داده‌ها اعمال می‌شوند، به ما می‌دهد. با استفاده از این الگوریتم، می‌توانیم روی مسائل پیچیده‌تری مانند تشخیص چهره کار کنیم.

کمبود داده یکی از مشکلاتی است که مانع استفاده از CNN می‌شود. با وجود اینکه می‌توان شبکه‌ها را با تعداد داده نسبتاً کمی، تقریباً ۱۰۰۰۰، آموزش داد، هرچه اطلاعات بیشتری در دسترس باشد، CNN بهتر تنظیم می‌شود. داده‌ها باید بدون‌نقص و دارای برچسب باشند تا CNN بتواند از آن‌ها استفاده کند و این چیزی است که باعث می‌شود کار کردن با آن‌ها زمان‌بر و نیازمند منابع سنگین محاسباتی باشد.

شبکه‌های عصبی پیچشی براساس یافته‌های علوم اعصاب [1] عمل می‌کنند. آن‌ها از لایه‌هایی از نورون‌های مصنوعی به نام نود [2] ساخته شده‌اند. این نودها توابعی هستند که مجموع وزنی ورودی‌ها را محاسبه می‌کنند و یک نگاشت فعال‌سازی [3] را برمی‌گردانند. این بخشِ پیچشی [4] شبکه عصبی [5] است.



شکل ۵.۲: بخش پیچشی شبکه عصبی

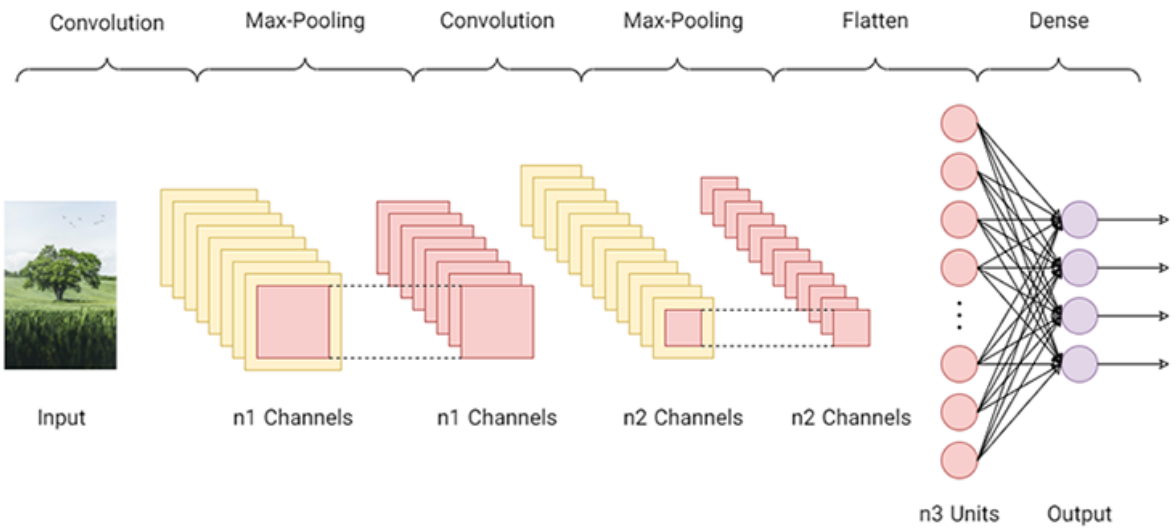
هر نود در یک لایه توسط مقادیر وزنی آن تعریف می‌شود. وقتی به یک لایه داده‌هایی را می‌دهید، برای مثال یک تصویر، مقادیر پیکسل را می‌گیرد و برخی از ویژگی‌های بصری را جدا می‌کند.

هنگامی که داده‌ها را به CNN می‌دهید، هر لایه نگاشت‌های فعال‌سازی را برمی‌گرداند. این نگاشت‌ها ویژگی‌های مهم مجموعه داده را شناسایی می‌کنند. اگر به CNN تصویری را بدهید، ویژگی‌های مبتنی بر مقادیر پیکسل مانند رنگ‌ها را شناسایی می‌کند و تابع فعال‌سازی را به شما ارائه می‌دهد.

معمولاً در تصاویر، CNN در ابتدا لبه‌های تصویر را پیدا می‌کند. سپس این تعریفِ جزئی از تصویر به لایه بعدی منتقل می‌شود و آن لایه شروع به شناسایی مواردی مانند گوشه‌ها و گروه‌های رنگی می‌کند. سپس این تعریف جدید از تصویر به لایه بعدی منتقل می‌شود و چرخه تا پیش‌بینی ادامه پیدا می‌کند.

همان طور که در تصویر زیر مشخص است، با افزایش تعداد لایه‌ها حداکثر تجمع (max-pooling) باید انجام شود. حداکثر تجمع فقط مرتبط‌ترین ویژگی‌ها از لایه موجود در نقشه فعال‌سازی را برمی‌گرداند و به لایه‌های بعدی منتقل می‌کند تا زمانی که به لایه آخر برسید. آخرین لایه CNN لایه طبقه‌بندی است که مقدار پیش‌بینی‌شده را براساس نگاشت فعال‌سازی تعیین می‌کند. اگر یک نمونه دست‌خط را به CNN بدهید، لایه‌ی طبقه‌بندی حروف موجود در تصویر را به شما می‌گوید. این همان چیزی است که وسایل نقلیه خودران برای تعیین اینکه یک شیء اتومبیل، شخص و یا یک مانع است، استفاده می‌کنند.

[1] neuroscience
[2] node
[3] activation map
[4] Convolutional
[5] Neural Network



شکل ۶.۲: معماری و لایه‌های یک شکبه عصبی پیچشی

انواع شبکه عصبی پیچشی به شرح زیر است:

- CNN یک‌بعدی: در این حالت، کرنل CNN در یک جهت حرکت می‌کند. CNN های یک‌بعدی معمولاً روی داده‌های سری زمانی استفاده می‌شوند.
- CNN دوبعدی: در این نوع از CNN ، کرنل‌ها در دو جهت حرکت می‌کنند. CNN های دوبعدی در برچسب‌گذاری و پردازش تصویر کاربرد دارند.
- CNN سه‌بعدی: این نوع CNN دارای کرنلی است که در سه جهت حرکت می‌کند. محققان از این نوع CNN در تصاویر سه‌بعدی مانند سی‌تی‌اسکن و MRI استفاده می‌کنند.

از آنجایی که بیشتر مسائل با داده‌های تصویر مرتبط هستند، اغلب از CNN های دوبعدی استفاده می‌شود. در ادامه برخی از کاربردهایی که ممکن است از CNN ها استفاده شود، آورده شده است.

- تشخیص تصاویر با پیش‌پردازش کم
- تشخیص دست‌خط‌های مختلف
- کاربردهای بینایی کامپیوتر (Computer Vision)
- استفاده در بانک‌داری بای خواندن ارقام در چک
- استفاده در سرویس‌های پستی برای خواندن کدپستی روی پاکت نامه

فصل ۳

کار مربوطه

۱.۳ استخراج ویژگی‌ها

استخراج ویژگی ها بخش بسیار مهمی در تجزیه و تحلیل و یافتن روابط بین چیزهای مختلف است. همانطور که قبلاً می دانیم که داده های ارائه شده از صدا را نمی توان مستقیماً توسط مدل ها درک کرد، بنابراین ما باید آنها را به یک قالب قابل درک تبدیل کنیم که استخراج ویژگی برای آن استفاده می شود. سیگنال صوتی یک سیگنال سه بعدی است که در آن سه محور زمان، دامنه‌ی نوسان و فرکانس را نشان می دهد.
هر فایل موسیقی اساساً از دو چیز مهم تشکیل شده است:

- نرخ نمونه [۱]
- داده نمونه [۲]

اکنون با کمک نرخ نمونه و داده های نمونه می توان چندین تغییر شکل روی آن انجام داد تا ویژگی های ارزشمندی را از آن استخراج کرد که در قسمت زیر آمده است:

۱. نرخ عبور از صفر[۳]: نرخی است که در آن یک سیگنال از مثبت به صفر به منفی یا از منفی به صفر به مثبت تغییر می کند.

$$zcr = \frac{1}{T-1}\sum_{t=1}^{T-1} 1_{\mathbb{R}<0}(s_t s_{t-1}) \tag{۱.۳}$$

۲. انرژی:

$$\mathbb{E}_s = \langle f(t), f(t)\rangle = \int_{-\infty}^{\infty} |f(t)|^2 dt \tag{۲.۳}$$

۳. آنتروپی: معیاری عددی برای اندازه گرفتن اطلاعات، یا تصادفی بودن یک متغیر تصادفی است. به بیان دقیق‌تر، آنتروپی یک متغیر تصادفی، متوسط اطلاعات آن است. با داشتن یک متغیر تصادفی گسسته X که مقادیری از الفبای X، آنتروپی برای آن به صورت زیر تعریف می‌شود:

$$H(x) = -\sum_{x\in X} p(x)\log p(x) \tag{۳.۳}$$

۴. مرکز طیفی: نشان می دهد که مرکز جرم طیف در کجا قرار دارد. از نظر ادراکی، ارتباط قوی با تأثیر روشنایی صدا دارد. گاهی به آن مرکز جرم طیفی نیز می گویند.

$$C_i = \frac{\sum_{k=1}^{Wf_L} kX_i(k)}{\sum_{k=1}^{Wf_L} X_i(k)} \tag{۴.۳}$$

۵. گسترش طیفی: دومین لحظه مرکزی طیف است که برای محاسبه آن باید انحراف طیف را از مرکز طیفی مطابق معادله زیر گرفت:

$$S_i = \sqrt{\frac{\sum_{k=1}^{Wf_L}(k-C_i)^2 X_i(k)}{\sum_{k=1}^{Wf_L} X_i(k)}} \tag{۵.۳}$$

[1] Sample Rate
[2] Sample Data
[3] Zero-crossing rate

۶. آنتروپی طیفی: محاسبه توزیع توان طیفی همراه با قابلیت پیش‌بینی سیگنال سری زمانی است. این آنتروپی بر اساس شانون و آنتروپی اطلاعات در داده های اطلاعاتی است. آنتروپی طیفی سیگنال توسط:

$$SE(F) = -\frac{1}{\log N_u \sum_u (p_u(F) \log_e P_u(F))} \tag{۶.۳}$$

$$SSH(F) = -\sum_u (P_h(F) \log_e P_h(F)), \tag{۷.۳}$$

به صورتی که $P_u(F)$ نشان دهنده‌ی تابع چگالی طیفی توان، $P_h(F)$ نشان دهنده تخمین آنتروپی شانون $(SSH(F))$، و N_u کل فرکانس‌ها را نشان می‌دهد.

۷. شار طیفی: شار طیفی تغییر طیفی بین دو فریم متوالی را اندازه گیری می کند و به عنوان اختلاف مجذور بین مقادیر نرمال شده طیف دو پنجره کوتاه مدت متوالی محاسبه می شود:

$$Fl_{(i,i-1)} = \sum_{k=1}^{Wf_L} (EN_i(k) - EN_{i-1}(k))^2, \tag{۸.۳}$$

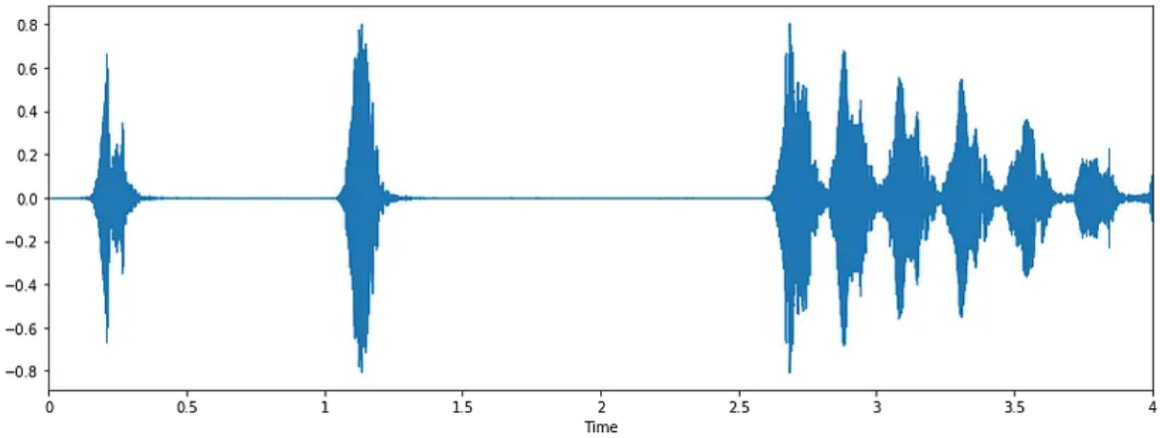
به صورتی که $EN_i(k) = \frac{X_i(k)}{\sum_{l=1}^{Wf_L} X_i(l)}$ نشان دهنده‌ی kامین ضریب نرمال شده در فریم iام است.

۸. افت تدریجیه طیفی: [1] نقطه ای است که درصد معینی از کل انرژی طیفی زیر آن قرار دارد. به تمایز بین صداهای گفتاری صدادار و بدون صدا کمک می کند.

۹. ضرایب طیفی فرکانس‌های ملتویی:[2] یک روش مهم در پردازش سیگنال‌های صوتی است. این روش با استفاده از تقسیم سیگنال صوتی به قطعات کوتاه و استخراج ویژگی‌های فرکانسی و زمانی از آن، در تشخیص و شناسایی گفتار و الگوهای صوتی مؤثر استفاده می‌شود. MFCC اطلاعات حیاتی درباره‌ی طیف فرکانسی و ویژگی‌های زمانی سیگنال صوتی را ارائه می‌دهد و در سیستم‌های تشخیص گفتاری و شناسایی الگوهای صوتی به کار می‌رود.

۱.۱.۳ Waveplots

صدای خام را می توان به عنوان یک نمودار موج تجسم کرد که نمونه ای از آن در شکل زیر داده شده است. یک نمودار موج، پوشش دامنه سیگنال را در برابر زمان ترسیم می کند. تجسم اینکه یک سیگنال چگونه به نظر می رسد می تواند مفید باشد، اما معمولاً برای مدل های یادگیری ماشین در پیش بینی ها مفید نیست. برای اینکه یک سیگنال مفید باشد، لازم است ویژگی های کمتر آشکار استخراج شود. معمولاً ویژگی‌های استخراج‌شده را می‌توان به دو دسته تقسیم کرد: زمانی که به ویژگی‌های وابسته به زمان می‌پردازد و طیفی که با ویژگی‌های وابسته به فرکانس سروکار دارد.

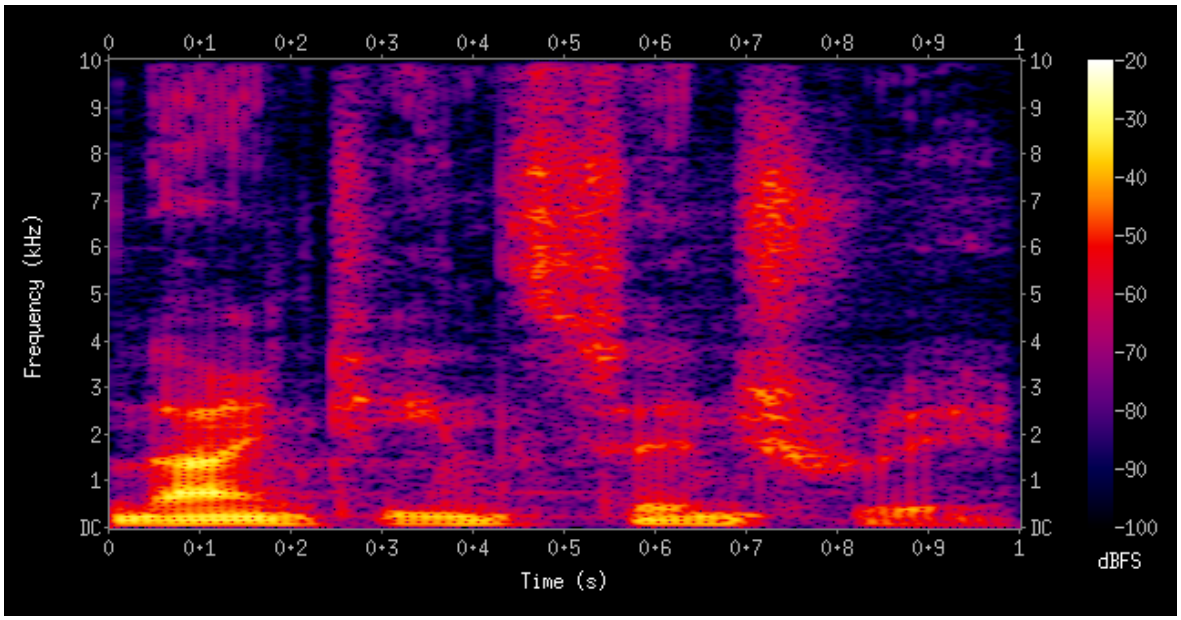


شکل ۱.۳: طرحی از یک Waveplot

[1] Spectral Roll off
[2] MFCC

۲.۱.۳ طیف نگاری

یک طرح طیف نگاره [1] روشی بصری برای نمایش قدرت سیگنال یا بلندی سیگنال در طول زمان در فرکانس‌های مختلف موجود در یک شکل موج خاص است. این به ما کمک میکند که در طول زمان در فرکانس‌های مختلف مشاهده کنیم چگونه سطوح انرژی تغییر می‌کند. طیف نگاره‌ها نمودارهای دو بعدی هستند که بعد سوم با رنگ‌ها نشان داده می‌شود. زمان از چپ (قدیمی‌ترین) به راست (جوان‌ترین) در امتداد محور افقی می‌گذرد. محور عمودی نشان‌دهنده فرکانس است که می‌توان آن را به صورت زیر و بم یا تن نیز در نظر گرفت، با کمترین فرکانس در قسمت پایین و بیشترین فرکانس در قسمت بالایی قرار گرفته است. دامنه (یا انرژی یا "بلندی") یک فرکانس خاص در یک زمان خاص با بعد سوم، رنگ، با آبی تیره مربوط به دامنه های کم و رنگ های روشن تر تا قرمز مربوط به دامنه های به تدریج قوی تر (یا بلندتر) نشان داده می شود.



شکل ۲.۳: طرحی از یک طرح نگاره

۲.۳ مجموعه داده

ابتدا برای اینکه یک مدل تشخیص دهنده‌ای را ایجاد کنیم نیاز است که مجموعه داده صوتی احساسی را استفاده کنیم، اگر چه برخی از این پایگاه‌های داده به دلیل حفظ حریم شخصی افراد در دسترس عموم توسعه دهنده‌گان این حوزه قرار نگرفته است، برخی از این پایگاه‌های داده مورد استفاده در مدل مولد و مدل تشخیص احساسات در این پروژه به شرح زیر است:

- Ravdess: این پایگاه داده شامل ۱۴۴۰ فایل است که هر گوینده دارای ۶۰ فایل صوتی است. این دیتاست شامل ۲۴ گوینده (۱۲ مرد و ۱۲ زن) است، که به زبان انگلیسی با گویش آمریکای شمالی صحبت کرده اند. احساسات مورد استفاده شامل موارد زیر است:
  - آرام (Calm)
  - شاد (Happy)
  - غمگین (Sad)
  - خشمگین (Angry)
  - ترسناک (Fearful)
  - حیرت‌زده (Surprise)
  - منزجر (Disgust)

  هر یک از موارد بالا در سه درجه عادی، خنثی، و شدید جمع‌آوری شده است.

[1] Spectogram

SAVEE ۱.۲.۳

پایگاه داده SAVEE از چهار مرد بومی زبان انگلیسی (مشخص شده به عنوان DC، JE، JK، KL)، دانشجویان کارشناسی ارشد و محققان دانشگاه ساری در سنین ۲۷ تا ۳۱ سال ثبت شد. عاطفه از نظر روانشناختی در دسته بندی های مجزا توصیف شده است: خشم، انزجار، ترس، شادی، اندوه و تعجب. یک دسته خنثی نیز اضافه شده است تا ضبط ۷ دسته احساسات را ارائه دهد.

محتوای متن شامل ۱۵ جمله TIMIT در هر احساس بود: ۳ جمله رایج، ۲ جمله خاص هیجان و ۱۰ جمله کلی که برای هر احساس متفاوت و از نظر آوایی متعادل بودند. ۳ جمله رایج و ۲ × ۶ = ۱۲ جمله خاص هیجانی به عنوان خنثی ثبت شد تا ۳۰ جمله خنثی ارائه شود. این منجر به ۱۲۰ بیان برای هر گوینده شد.

ESD ۲.۲.۳

پایگاه داده ESD توسط دانشگاه ملی سنگاپور (NUS) و دانشگاه فناوری و طراحی سنگاپور (SUTD) در دسترس است. پایگاه داده ESD شامل ۳۵۰ گفتار موازی است که توسط ۱۰ انگلیسی بومی و ۱۰ فرد چینی (Mandarian) صحبت می شود و ۵ کلاس احساسی (خنثی، شادی، خشم، غم و حیرت) را پوشش می دهد. بیش از ۲۹ ساعت داده گفتار در محیط آکوستیک کنترل شده ثبت شده است. بنابراین، برای مطالعات تبدیل صدای عاطفی چند گوینده و چند زبانه مناسب است.

کاربردهای این مجموعه داده به صورت زیر است:

- تبدیل صدای احساسی (تک زبانه و چند زبانه، وابسته به گوینده و مستقل از گوینده)
- تبدیل صدا (تک زبانه و چند زبانه)
- متن به گفتار احساسی
- بیان متن به گفتار

Parameter	Mandarin						English					
	Neu	Ang	Sad	Hap	Sur	All	Neu	Ang	Sad	Hap	Sur	All
# speakers	10	10	10	10	10	10	10	10	10	10	10	10
# utterances per speaker	350	350	350	350	350	1,750	350	350	350	350	350	1,750
# unique utterances	350	350	350	350	350	350	350	350	350	350	350	350
# characters/words per speaker	4,005	4,005	4,005	4,005	4,005	20,025	2,203	2,203	2,203	2,203	2,203	11,015
# unique characters/words	939	939	939	939	939	939	997	997	997	997	997	997
Avg. utterance duration [s]	3.23	2.68	4.04	2.84	3.32	3.22	2.61	2.80	2.98	2.70	2.73	2.76
Avg. character/word duration [s]	0.28	0.23	0.35	0.25	0.29	0.28	0.41	0.44	0.47	0.43	0.43	0.44
Total duration [s]	11,305	9,380	14,140	9,940	11,620	56,385	9,135	9,800	10,430	9,450	9,555	48,370

Emotion abbreviations are used as follows: *Neu* stands for neutral, *Ang* stands for anger, *Sad* stands for sadness, *Hap* stands for happiness and *Sur* stands for surprise. The number of characters is reported for Mandarin, and the number of words is reported for English.

شکل ۳.۳: جزئیات مجموعه داده ESD

Tess ۳.۲.۳

در این مجموعه از ۲۰۰ کلمه هدف در عبارت حامل "کلمه را بگویید" توسط دو بازیگر زن (۲۶ و ۶۴ ساله) بیان شده است و ضبط هایی از مجموعه انجام شده است که هر یک از هفت احساس (خشم، انزجار، ترس، شادی، غافلگیری دلپذیر، غمگینی و خنثی) را به تصویر می کشد. در مجموع ۲۸۰۰ نقطه داده (فایل صوتی) وجود دارد. مجموعه داده به گونه ای سازماندهی شده است که هر یک از دو بازیگر زن و احساسات آنها در پوشه مخصوص به خود قرار دارند. در آن، تمام ۲۰۰ کلمه هدف فایل صوتی را می توان یافت. فرمت فایل صوتی فرمت WAV می باشد.

CREMA-D ۴.۲.۳

CREMA-D یک مجموعه داده عاطفی بازیگر چندوجهی از ۷۴۴۲ کلیپ اصلی از ۹۱ بازیگر است. این کلیپ ها از ۴۸ بازیگر مرد و ۴۳ بازیگر زن بین ۲۰ تا ۷۴ سال بود که از نژادها و قومیت های مختلف (آمریکای آفریقایی، آسیایی، قفقازی، اسپانیایی تبار و سایر قومیتها) بودند.

بازیگران از مجموعه ای از ۱۲ جمله صحبت کردند. جملات با استفاده از یکی از شش احساس مختلف (خشم، انزجار، ترس، خوشحالی، خنثی، و غمگین) و چهار سطح هیجانی مختلف (کم، متوسط، زیاد و نامشخص) ارائه شدند. شرکت کنندگان بر اساس ارائه ترکیبی سمعی و بصری، ویدئو به تنهایی و صوت به تنهایی، احساسات و احساسات را ارزیابی کردند. با توجه به تعداد زیاد رتبه‌بندی‌های مورد نیاز، این تلاش به صورت جمعی انجام شد و در مجموع ۲۴۴۳ شرکت‌کننده هر کدام ۹۰ کلیپ منحصر به فرد، ۳۰ کلیپ صوتی، ۳۰ تصویری و ۳۰ کلیپ صوتی و تصویری را رتبه‌بندی کردند. ۹۵ درصد از کلیپ ها بیش از ۷ امتیاز دارند.