تولید موسیقی مبتنی بر پردازش زبان طبیعی

محمد امین کیانی[1] - ۴۰۰۳۶۱۳۰۵۲
ارائه‌ی درس مبانی پردازش زبان و گفتار - دکتر برادران
نیم‌سال دوم تحصیلی ۴۰۴-۴۰۳

[1] دانشجوی کارشناسی، دانشکده‌ی کامپیوتر، دانشگاه اصفهان، اصفهان،
Aminkianiworkeng@gmail.com
https://github.com/M-Amin-Kiani/NLP-Proj

چکیده

در سال‌های اخیر پیشرفت‌های قابل‌توجهی در زمینه‌ی تولید موسیقی از متن با به‌کارگیری پردازش زبان طبیعی (NLP) و مدل‌های یادگیری عمیق صورت گرفته است. این حوزه که متن را به عنوان ورودی گرفته و قطعه‌ی موسیقی متناظر با محتوای احساسی و معنایی آن تولید می‌کند، ترکیبی از تحلیل احساسات متن و تولید خودکار موسیقی است. در این مقاله مروری، جدیدترین پژوهش‌ها و مدل‌های پنج سال اخیر در زمینه‌ی تولید موسیقی مبتنی بر متن را به‌صورت جامع بررسی می‌کنیم. ابتدا کارهای مرتبط شامل روش‌های کلاسیک مبتنی بر قواعد (مانند نگاشت احساسات متن به ویژگی‌های موسیقی) و نسل اول مدل‌های یادگیری ماشین معرفی می‌شوند. سپس مدل‌های نوین و برجسته نظیر MusicLM، Riffusion، AudioGen، MusicGen و سایر روش‌های مطرح همراه با جزئیات معماری و فرمول‌بندی آن‌ها تشریح می‌گردند. این مدل‌ها بر اساس تکنیک یادگیری ماشین (ترنسفورمر، انتشار(diffusion)، خودبازگشتی و ...)، نوع ورودی (متن خام، متن همراه با برچسب احساسی یا ملودی اولیه) و نوع خروجی (نمادگذاری MIDI یا موج صوتی) طبقه‌بندی و مقایسه شده‌اند. مزایا، چالش‌ها و محدودیت‌های هر دسته (از جمله کیفیت صوتی، همخوانی با متن، نیاز به داده‌ی آموزشی گسترده و مسائل حق‌کپی) به تفصیل بحث می‌شود. در ادامه، یک پیاده‌سازی نمونه‌ی کد متن‌باز برای تبدیل متن به موسیقی (با استفاده از کتابخانه‌های موجود) ارائه شده است تا درک عملی این فناوری میسر گردد. نتایج بررسی نشان می‌دهد که علی‌رغم پیشرفت‌ها (مانند تولید موسیقی با کیفیت بالا از توضیحات متنی پیچیده)، همچنان چالش‌هایی نظیر کمبود داده‌های برچسب‌خورده‌ی متن-موسیقی، محدودیت در بیان احساسات بلندمدت و ارزیابی کیفی مدل‌ها پابرجاست. در پایان، ضمن جمع‌بندی، به جهت‌گیری‌های پژوهشی آینده همچون بهبود هم‌آمیزی متن و موسیقی، افزایش قابلیت کنترل‌پذیری توسط کاربر، و استفاده از مدل‌های زبانی بزرگ برای ارتقای فهم و تولید خلاقانه‌تر موسیقی اشاره شده است.

کلمات کلیدی

تولید موسیقی از متن، تحلیل احساسات، پردازش زبان طبیعی، مدل‌های مولد، ترنسفورمر، مدل انتشار، MusicLM، MusicGen، Riffusion.

۱- مقدمه

تلفیق موسیقی و احساسات از دیرباز مورد توجه حوزه‌های مختلف هنر و فناوری بوده است. موسیقی به عنوان "زبان جهانی" نقشی کلیدی در بیان احساسات انسانی دارد و همواره با ادبیات و متن در ارتباط بوده است. در سال‌های اخیر، با پیشرفت الگوریتم‌های هوش مصنوعی، ایده‌ی تولید خودکار موسیقی بر پایه‌ی متن مورد توجه قرار گرفته است. منظور از تولید موسیقی از متن، فرایندی است که در آن یک سیستم هوشمند با دریافت ورودی متنی (مانند توضیحاتی در مورد حال‌وهوا، سبک و محتوای احساسی)، یک قطعه موسیقی منطبق با آن متن خلق می‌کند. این مسئله ماهیتی چندرسانه‌ای دارد و از دو حوزه‌ی عمده بهره می‌گیرد: (۱) پردازش زبان طبیعی برای درک معنایی و احساسی متن، و (۲) سیستم‌های تولید موسیقی برای ساخت ملودی،

هارمونی و تنظیم موسیقایی متناسب با آن. ترکیب این دو حوزه امکان می‌دهد احساسات مستتر در کلمات به زبان موسیقی ترجمه شود.

اهمیت موضوع: تولید موسیقی به کمک رایانه پیشینه‌ای چند ده‌ساله دارد. روش‌های اولیه عموماً بر پایه‌ی دانش صریح موسیقی (نظریه‌های موسیقی) و قواعد از پیش تعریف‌شده بودند. به بیان دیگر، سیستم‌های قدیمی‌تر از رویکردهای مبتنی بر دانش استفاده می‌کردند که در آن‌ها آهنگسازی خودکار بر اساس قوانین موسیقی (نظیر قواعد هارمونی کلاسیک) یا نگاشت‌های از پیش تعیین‌شده انجام می‌شد. به عنوان مثال، الگوریتم‌های تکاملی و جست‌وجوی فضای نت‌ها به کمک قواعد موسیقایی از چند دهه پیش برای آهنگسازی به کار رفته‌اند. اما این رویکردها غالباً انعطاف‌پذیری و تنوع محدودی داشتند زیرا خلاقیت آن‌ها مستقیماً وابسته به قواعد صریح برنامه‌ریزی‌شده بود. در مقابل، پیشرفت یادگیری عمیق در سال‌های اخیر زمینه‌ساز ظهور روش‌های داده–محور در تولید موسیقی شده است. امروزه مدل‌های مولد عمیق می‌توانند با یادگیری از مجموعه‌های بزرگ داده‌ی موسیقی، الگوهای پیچیده‌ای را فرابگیرند که دستیابی به آن‌ها صرفاً با قواعد صریح ممکن نبود. در حوزه‌ی متن–به–موسیقی نیز رویکردهای یادگیری ماشین امکان استفاده از داده‌های همزمان متن و موسیقی را فراهم کرده‌اند تا ارتباط میان توصیفات زبانی و ویژگی‌های صوتی موسیقی آموخته شود. این تحول، تمرکز پژوهش‌ها را از طراحی دستی قواعد به سمت طراحی مدل‌هایی معطوف کرده که با یادگیری از داده‌های حجیم، دانش ضمنی آهنگسازی را کسب می‌کنند.

چالش اتصال متن و موسیقی: با وجود پیشرفت‌ها، یکی از چالش‌های اساسی این حوزه نحوه‌ی هم‌ترازی و تطبیق متن ورودی با ساختار موسیقی خروجی است. متن دارای ویژگی‌هایی مانند دستور زبان، معنا و احساسات است در حالی که موسیقی ساختاری زمانی، شامل نُت‌ها، آکوردها، تمپو، دینامیک و تُنالیته دارد. ارتباط دادن مستقیم این دو ساختار ناهمگون آسان نیست. برای مثال، توصیف متنی «یک ملودی شاد با ویولن در پس‌زمینه‌ی ریتم تند گیتار برقی» باید به رویدادهای دقیق موسیقایی (نظیر گام ماژور، ساز ویولن با ملودی بالا، وجود گیتار برقی با ریتم تند و میزان سرعت مشخص) ترجمه شود. مدل هوش مصنوعی باید درک زبانی کافی از متن داشته باشد (مثلاً بداند «شاد» به گام ماژور یا ملودی‌های زیر مرتبط است) و همزمان دانش موسیقایی ضمنی برای پیاده‌سازی آن مفاهیم در صوت را کسب کرده باشد. علاوه بر این، موسیقی دارای بُعد زمانی است و نیاز به انسجام بلندمدت دارد؛ مدل باید بتواند یک قطعه‌ی گوش‌نواز و منسجم بسازد که صرفاً توالی تصادفی صداها نباشد.

پردازش احساسات متن: یکی از جنبه‌های کلیدی در تبدیل متن به موسیقی، استخراج و بهره‌گیری از محتوای احساسی متن (sentiment/emotion) است. معمولاً متن ورودی ممکن است حاوی حالات احساسی (شاد، غمگین، هیجان‌انگیز، آرامش‌بخش و ...) باشد یا به طور ضمنی فضایی احساسی را القا کند. تبدیل این احساس به المان‌های موسیقی یکی از اهداف مهم این حوزه است. برای نمونه، یک متن با لحن مثبت و شادی‌بخش احتمالاً منجر به موسیقی در گام ماژور، با سرعت نسبتاً بالا و سازبندی روشن (مثلاً فلوت یا ویولن) می‌شود؛ برعکس، متن غمگین یا منفی ممکن است موسیقی در گام مینور با تمپوی آهسته و سازهایی با تُن تیره (مانند ویولن‌سل یا پیانو در اکتاو پایین) را بطلبد. تحلیل احساسات شاخه‌ای از NLP است که به تشخیص و رده‌بندی احساس متن (مثبت، منفی، خنثی یا طبقه‌بندی‌های غنی‌تر مانند هیجان، اندوه، خشم و غیره) می‌پردازد. ترکیب این تحلیل با سیستم‌های تولید موسیقی امکان کنترل احساسی خروجی را فراهم می‌کند. کاربردهای سیستم‌های متن–به–موسیقی گسترده است: از آهنگسازی پس‌زمینه برای فیلم‌ها و بازی‌های ویدئویی بر اساس توصیف صحنه یا داستان، تا کتاب‌های صوتی تعاملی که متناسب با حال‌وهوای هر فصل کتاب موسیقی متن خلق می‌کنند، و حتی کمک به موسیقی‌درمانی که در آن بر اساس بیان کلامی احساسات بیمار، موسیقی آرام‌بخش یا نشاط‌آور تولید شود [۱].

مثال‌های جالب دیگری نیز پیشنهاد شده است، از جمله تولید موسیقی همراه با تحلیل احساسات شبکه‌های اجتماعی (مثلاً پخش موسیقی منعکس‌کننده‌ی احساس عمومی توییت‌های یک منطقه جغرافیایی). این کاربردها نشان می‌دهد سیستم‌های متن–به–موسیقی می‌توانند نقش یک پل احساسی را ایفا کنند که داده‌های متنی را به تجربه‌ی شنیداری تبدیل می‌کند. با توجه به سرعت پیشرفت این حوزه، نیاز به یک مطالعه مروری جامع احساس می‌شود تا پژوهشگران دیدی منسجم نسبت به روش‌های مختلف و وضعیت فعلی تکنولوژی پیدا کنند. در ادامه این مقاله، ابتدا کارهای انجام‌شده‌ی پیشین و زمینه‌ساز معرفی می‌گردند. سپس به تفصیل مدل‌های جدیدتر (بالاخص سال‌های ۲۰۱۹ تا ۲۰۲۴) که متن را به موسیقی تبدیل می‌کنند، بررسی خواهد شد. در هر بخش تلاش شده‌است فرمولاسیون ریاضی یا معماری مدل‌ها نیز (در حد امکان) ارائه شود تا درک فنی دقیق‌تری حاصل گردد. همچنین یک چارچوب طبقه‌بندی برای مقایسه‌ی روش‌ها از منظر نوع ورودی/خروجی و تکنیک یادگیری ارائه می‌کنیم. نهایتاً، ضمن بحث در مورد چالش‌ها و روندهای آینده، یک نمونه کد پیاده‌سازی متن‌باز برای حل مسئله‌ی متن–به–موسیقی معرفی خواهد شد تا جنبه‌های عملی موضوع روشن‌تر شود.

۲- کارهای مرتبط و سوابق پژوهشی

در این بخش مروری بر مهم‌ترین پژوهش‌های مرتبط با تولید موسیقی از متن انجام می‌شود. ابتدا رویکردهای مبتنی بر نگاشت احساسی قوانین‌محور و روش‌های سنتی معرفی می‌شوند و سپس گذار به روش‌های یادگیری عمیق و مدل‌های مدرن‌تر تشریح می‌گردد.

۲-۱- روش‌های مبتنی بر قوانین و عاطفه‌کاوی کلاسیک:

نخستین تلاش‌ها برای تولید موسیقی از روی متن عمدتاً بر پایه‌ی روش‌های قواعد–محور و تکیه بر دانش مهندسی‌شده بودند. یک نمونه شاخص، سیستم TransProse (دیویس و محمد، ۲۰۱۴) است که با هدف ایجاد موسیقی از روی ادبیات داستانی (رُمان) طراحی شد [۲]. در TransProse، متن رمان از منظر فراوانی کلمات احساسی تحلیل می‌شود و سپس با استفاده از نگاشت‌های از پیش تعیین‌شده بین مفاهیم احساسی و عناصر موسیقی، یک قطعه پیانوی ساده تولید می‌گردد. به طور مثال، این سیستم چگالی واژگان احساسی در هر بخش متن را محاسبه کرده و بر اساس آن ویژگی‌هایی نظیر تمپو (سرعت) و گام (ماژور برای احساس مثبت، مینور برای منفی) را تعیین می‌کند. همچنین مجموعه‌ای از قواعد برای انتخاب توالی نُت‌ها متناسب با تغییرات احساس در طول داستان به کار رفته است.

هرچند رویکرد TransProse کاملاً ساده‌انگارانه است (چون فقط بر چند ویژگی کلی تمرکز دارد)، اما نمایانگر اولین گام‌ها در تبدیل مستقیم متن ادبی به موسیقی احساسی بود. این کار نشان داد که می‌توان از طریق قواعد ثابت، پل اولیه‌ای بین متن و موسیقی برقرار کرد. پس از آن، پژوهش‌های قواعد–محور دیگری نیز انجام شد؛ برای مثال، ویلیامز و همکاران (۲۰۱۵) سیستمی برای تولید موسیقی متن بازی‌های رایانه‌ای ارائه کردند که گراف صحنه‌های بازی را با برچسب‌های احساسی مشخص می‌کرد و سپس یک مدل مارکوف برای ساخت ملودی مطابق آن احساسات به کار می‌برد [۳]. در پژوهشی دیگر، دیویس و محمد (همان TransProse) تلاش مشابهی را برای رمان‌های کلاسیک انجام دادند و با استفاده از روش‌های مبتنی بر فرهنگ لغت احساس (lexicon-based) احساسات متن را استخراج و با قواعد موسیقایی ترکیب کردند. به طور کلی، در این نسل از کارها، مراحل حل مسئله تفکیک‌شده بود: ابتدا تحلیل احساس متن به‌صورت نمادین (مثلاً تشخیص اینکه پاراگرافی غمگین یا شاد است) و سپس تولید موسیقی بر پایه قوانین مهندسی‌شده برای آن احساس. مزیت اصلی روش‌های قواعد–محور، سادگی و شفافیت آنها است. طراح سیستم دقیقاً می‌داند که چرا یک خروجی موسیقی خاص تولید شده (زیرا مستقیماً توسط قوانین کدنویسی‌شده تعیین شده است). اما ضعف بزرگشان عدم انعطاف و فقدان خلاقیت است. موسیقی تولیدشده توسط این روش‌ها اغلب یکنواخت و مکانیکی به نظر می‌رسد و تنوع یا پیچیدگی زیادی ندارد. به عنوان مثال، اگر قاعده این باشد که متن غمگین یعنی گام مینور و سرعت آهسته، تمام خروجی‌های غمگین به هم شبیه خواهند بود. همچنین این سیستم‌ها قادر به درک عمیق متن یا ترکیب ظرائف چندگانه (احساسات مرکب، تغییرات تدریجی حالت و ...) نیستند.

۲-۲- یادگیری عمیق برای تولید موسیقی با احساس(نسل اول):

با پیشرفت شبکه‌های عصبی عمیق در دهه ۲۰۱۰، پژوهشگران شروع به به‌کارگیری آن‌ها در آهنگسازی خودکار کردند. مدل‌های مشهوری نظیر DeepBach [۱] و MusicVAE در همین دوره توانستند قطعات چندصدایی یا ملودی‌های متنوعی تولید کنند، اما این مدل‌ها عمدتاً کنترل‌ناپذیر بودند؛ به این معنا که کاربر به سختی می‌توانست احساس یا سبک خروجی را تعیین کند. به عنوان مثال، اگر یک شبکه‌ی عصبی روی مجموعه‌ای از موسیقی کلاسیک آموزش دیده بود، خروجی آن همیشه رنگ‌وبویی شبیه همان داده‌ها داشت و تغییر دلخواه در احساس قطعه (مثلاً شادتر یا غمگین‌تر کردن) آسان نبود [۱]. برای حل این مشکل، پژوهش‌هایی به سمت مهار یا کنترل احساس در مدل‌های مولد پیش رفتند. یکی از اولین کارهای عمیق در این زمینه توسط فررا و وایت‌هد (۲۰۱۹) ارائه شد که عنوان مقاله آن‌ها «یادگیری تولید موسیقی با احساس» بود [۴]. آن‌ها یک مدل مبتنی بر شبکه عصبی LSTM چندلایه معرفی کردند که می‌توانست موسیقی نمادین (MIDI) چندصدایی تولید کند و به طور صریح توسط یک پارامتر احساس (مثبت یا منفی) هدایت شود. ایده‌ی کلیدی کار آن‌ها الهام‌گرفته از یک پژوهش در حوزه‌ی زبان توسط ردفورد و همکاران (۲۰۱۷) بود: در مدل زبانی GPT اولیه مشاهده شده بود که یک نرون منفرد در LSTM وجود دارد که نقش «نرون احساس» را بازی می‌کند و مقدار آن حس جمله (مثبت/منفی) را کدگذاری می‌کند. با دستکاری مقدار این نرون می‌شد جملات شاد یا غمگین تولید کرد. فررا و وایت‌هد همین ایده را به موسیقی تعمیم دادند. آن‌ها ابتدا یک مدل مولد LSTM را روی یک مجموعه داده از موسیقی‌های کلاسیک آموزش دادند (وظیفه: پیش‌بینی نت بعدی) و سپس نشان دادند این مدل به‌صورت خودبه‌خود ویژگی‌هایی در حالت نهان یاد گرفته که با احساس قطعه مرتبط است. برای به‌دست آوردن یک کنترل صریح، آن‌ها از یک دسته‌بند لجستیک استفاده کردند که بردار حالت نهان LSTM را به برچسب احساس (مثبت یا منفی) نگاشت می‌کرد. به این ترتیب، مدل قادر شد هنگام تولید موسیقی، با تنظیم جهت این بردار نهان به سمت مثبت یا منفی، خروجی شادتری یا غمگین‌تری ایجاد کند. برای آموزش و ارزیابی، آن‌ها یک مجموعه‌داده جدید به نام VGMIDI شامل ۸۲۳ قطعه موسیقی بازی‌های ویدئویی در فرمت MIDI ایجاد کردند که ۹۵ تای آن‌ها را ۳۰ نفر انسان از نظر ارزش (شاد/غمگین) و برانگیختگی برچسب‌گذاری احساسی کرده بودند. سپس مدل خود را روی این داده‌ها آزمودند. نتایج گزارش‌شده نشان داد که این مدل کنترل‌شونده از نظر تفکیک احساس توانست عملکرد بهتری از یک مدل LSTM معمولی (که به صورت نظارت‌شده برای طبقه‌بندی احساس آموزش دیده بود) داشته باشد. علاوه بر آن، در یک آزمایش شنیداری کاربر، قطعات تولیدی مدل با تنظیم "مثبت" توسط اکثر شنوندگان واقعاً شاد تلقی شدند؛ هرچند قطعات "منفی" همیشه کاملاً غمگین به نظر نمی‌رسیدند و بعضاً آمیخته‌ای از احساسات بودند. این یافته حاکی از آن بود که کنترل احساس در موسیقی ممکن است برای احساسات منفی سخت‌تر باشد (شاید به این دلیل که موسیقی غمگین می‌تواند تفاسیر متنوعی داشته باشد). به طور کلی، کار فررا و وایت‌هد نوآوری مهمی محسوب می‌شد زیرا برای اولین بار یک مدل یادگیری عمیق ارائه داد که تولید موسیقی با برون‌ده احساسی مشخص را میسر می‌ساخت. این کار مسیر را برای تحقیقات بعدی در زمینه‌ی ترکیب اطلاعات احساسی و مدل‌های مولد موسیقی هموار کرد. شایان ذکر است که در کنار این تلاش، سایر پژوهش‌ها نیز به سراغ کنترل‌پذیرکردن جنبه‌های دیگر موسیقی رفتند. برای مثال، مونتیث و همکاران (۲۰۱۲) از مدل‌های پنهان مارکوف (HMM) جداگانه برای هر کلاس احساس (شاد، غمگین، خشمگین و ... بر اساس یک مدل دسته‌بندی احساسی) بهره بردند تا ملودی‌هایی متناسب با هر طبقه ایجاد کنند. هرچند این رویکرد هنوز یادگیری عمیق نبود، اما نشانگر اهمیت کنترل عاطفه در آهنگسازی خودکار بود. همچنین اسکیریا و همکاران (۲۰۱۷) یک چارچوب به نام MetaCompose توسعه دادند که به‌صورت بلادرنگ برای بازی‌ها موسیقی تولید می‌کرد؛ آن‌ها از روش‌های تکاملی (مانند الگوریتم ژنتیک) برای خلق ملودی استفاده کردند و این ملودی را بر اساس برچسب احساسی صحنه‌ی بازی تنظیم می‌کردند. این رویکرد نیز نوعی کنترل احساسی اما با منطق تکاملی/ابتکاری بود.

۲-۳- مدل‌های یادگیری عمیق نسل جدید (متن به موسیقی):

از حوالی سال ۲۰۱۹ به بعد، شاهد ظهور مدل‌های بسیار بزرگتر و قدرتمندتری بوده‌ایم که مستقیماً قادر به تولید موج صوتی موسیقی از روی توضیحات متنی هستند. این مدل‌ها عموماً تحت تأثیر موفقیت‌های چشمگیر مدل‌های تولید تصویر از متن (نظیر DALL-E و Stable Diffusion) شکل گرفته‌اند. ایده‌ی کلی بسیاری از آن‌ها این است که ابتدا متن ورودی را به یک بردار یا توکن‌های نهان تبدیل می‌کنند (با استفاده از یک مدل زبان یا انکدر متنی)، سپس یک شبکه مولد آن بردار را به سیگنال صوتی تبدیل

می‌کند. در ادامه به برجسته‌ترین نمونه‌های این نسل و ویژگی‌های هرکدام می‌پردازیم.

۲-۳-۱- AudioGen (مدل مولد صوت متنی):

یکی از اولین کارهای پیشگام در تولید مستقیم صوت از متن، مدل AudioGen بود که توسط گروهی از پژوهشگران (Kreuk و همکاران) ارائه و در سال ۲۰۲۳ در ICLR منتشر شد [۵]. AudioGen در واقع برای تولید انواع صدا (شامل افکت‌های صوتی محیطی و نیز موسیقی ساده) از کپشن متنی طراحی شد. این مدل از رویکرد خودبازگشتی (autoregressive) بهره می‌گیرد؛ بدین صورت که ابتدا صوت را توسط یک کدک عصبی صوتی (مانند مدل EnCodec از متا) به توکن‌های گسسته تبدیل می‌کنند، سپس یک ترنسفورمرDecoder یاد می‌گیرد توالی این توکن‌ها را بر اساس متن تولید کند. در AudioGen برای تبدیل متن به بردار از یک انکدر متنی T5 (یک مدل زبانی ترنسفورمر بزرگ) استفاده شده است. متن پس از تبدیل به بردار، به بخش Decoder ترنسفورمر داده می‌شود تا خروجی صوتی (به شکل کدهای EnCodec) به صورت autoregressive پیش‌بینی شود. طول توکن‌های صوتی به علت نرخ نمونه‌برداری بالا (مثلاً ۲۴ کیلوهرتز) بسیار زیاد است و این چالشی برای مدل بود. طراحان AudioGen برای کاهش پیچیدگی محاسباتی از یک ترفند به نام مدل‌سازی چند-جریانی (multi-stream) استفاده کردند که عملاً کدهای صوتی را به چند توالی موازی کوتاه‌تر می‌شکند و مدل آن‌ها را همزمان تولید می‌کند. همچنین برای تقویت تطابق خروجی با متن از تکنیک راهنمایی بدون‌دسته‌بندی (Classifier-Free Guidance) استفاده شد، مشابه آنچه در مدل‌های دیفیوژن متن-تصویر انجام می‌شود. به زبان ساده، مدل گاهی خروجی را بدون شرط (متن) پیش‌بینی می‌کند و گاهی با شرط متن، و با ترکیب آن‌ها خروجی نهایی را هدایت می‌کند؛ این کار باعث می‌شود موسیقی تولیدشده بیشتر با توصیف متنی همخوان شود. AudioGen روی مجموعه‌ای متشکل از ۱۰ دیتاست مختلف صوتی-متنی (شامل AudioSet و AudioCaps و غیره) آموزش دید تا گستره‌ی متنوعی از صداها را یاد بگیرد. ارزیابی‌های گزارش‌شده نشان داد که AudioGen چه در معیارهای عینی (مثلاً فاصله فروشه یا سایر معیارهای کیفیت سیگنال) و چه در آزمون‌های ذهنی شنیداری، عملکرد بهتری نسبت به مدل‌های پایه قبلی داشته است. توسعه AudioGen از این جهت حائز اهمیت است که نشان داد مدل‌های ترنسفورمر بزرگ می‌توانند مستقیماً توالی‌های صوتی با کیفیت را از توضیحات متنی تولید کنند، به شرط آن‌که بازنمایی صوت به صورت مناسب (توسط کدک) انجام شود و معماری مدل برای غلبه بر طول زیاد توالی صوت طراحی گردد.

۲-۳-۲- DiffSound و AudioLDM (مدل‌های انتشار برای صوت):

تقریباً هم‌زمان با AudioGen، رویکرد دیگری مبتنی بر مدل‌های انتشار (Diffusion Models) برای تولید صوت از متن مطرح شد. یکی از آن‌ها مدل DiffSound بود [۱۳]. در DiffSound از یک انکدر CLIP (که برای تطبیق متن و تصویر آموزش دیده است) جهت استخراج ویژگی متنی استفاده شد و سپس یک شبکه انتشار وظیفه داشت اسپکتروگرام مل (تصویر زمان-فرکانس صوت) را بر اساس آن ویژگی متنی تولید کند. ایده این بود که با تولید طیف‌نگاشت و تبدیل آن به صوت (مثلاً توسط تبدیل معکوس فوریه)، می‌توان صدای متناظر را بدست آورد. مدل DiffSound نیز روی مجموعه داده‌هایی نظیر AudioCaps آموزش دید و با وجود حجم نسبتاً محدود داده‌های موازی (چند هزار ساعت)، نتایج امیدوارکننده‌ای در تولید افکت‌های صوتی مختلف به دست آورد. مدل دیگر AudioLDM نام دارد که مستقیماً یک چارچوب انتشار در فضای ویژگی نهفته را برای صوت ارائه کرد. AudioLDM در سال ۲۰۲۳ معرفی شد و ایده‌ی آن این است که یک مدل انتشار نهفته (مشابه latent diffusion در تصاویر) بسازیم که به جای پیکسل‌های تصویر، روی بردارهای ویژگی صوت (مثلاً خروجی یک autoencoder صوتی) کار کند. AudioLDM نیز از یک انکدر زبان (Flan-T5) برای فهم متن بهره می‌گیرد و سپس یک شبکه انتشار مشروط به متن نویز گاوسی را به ویژگی صوتی تبدیل می‌کند [۱۴]. این مدل نیز به دلیل کار در فضای نهفته، نسبت به طول سیگنال خام، کارآمدتر است و کیفیت صوتی خوبی ارائه داد. به دنبال AudioLDM، مدل‌های دیگری مانند Tango و Tango2 نیز مطرح شدند که با استفاده از مدل‌های زبانی بزرگ (LLM) جهت درک بهتر متن و سپس انتشار نهفته صوتی، سعی در بهبود کنترل‌پذیری داشتند [۶].

۲-۳-۳- Riffusion (کاربرد Stable Diffusion در موسیقی):

از مدل‌های بسیار مشهور سال‌های اخیر در میان عموم، Riffusion است [۷]. این مدل که اواخر ۲۰۲۲ معرفی شد، در واقع یک پروژه‌ی خلاقانه توسط دو توسعه‌دهنده (Forsgren و Martiros) بود که نشان داد می‌توان مدل انتشار متن-تصویر Stable Diffusion را با کمی دستکاری برای تولید موسیقی به کار گرفت. ایده Riffusion ساده اما جذاب است: ابتدا طیف‌نگاشت (spectrogram) یک قطعه موسیقی به‌صورت تصویر درنظر گرفته می‌شود، سپس Stable Diffusion را روی این فضای تصویری موسیقایی آموزش می‌دهند تا بتواند تصویر طیف‌نگاشت را بر اساس یک متن تولید کند. در واقع Riffusion یک مدل Stable Diffusion بود که به جای عکس‌های معمولی، با تصاویر طیف‌نگاشت آهنگ‌هاfine-tune شده بود. به این ترتیب، با دادن یک پرامپت متنی (مثلاً "آهنگ جاز با ساکسوفون آرام")، مدل تصویری از طیف‌نگاشت ایجاد می‌کند که وقتی توسط تبدیل فوریه معکوس به صوت تبدیل شود، صدایی شبیه موسیقی جاز با آن مشخصات خواهد داد. وجه تمایز Riffusion در تعاملی و بلادرنگ بودن آن است. این مدل قادر بود قطعات کوتاه موسیقی (چند ثانیه‌ای) را تقریباً به صورت آنی تولید کند و حتی به صورت پیوسته خروجی خود را حین تغییر متن، به‌روزرسانی نماید. به همین خاطر برای کاربردهای خلاقانه زنده (مثلاً اجرای موسیقی تعاملی بر اساس توضیحات آنی کاربر) مناسب بود. البته Riffusion محدودیت‌های محسوسی داشت: طول قطعات تولیدی کوتاه بود (معمولاً چند ثانیه) و ساختار بلندمدت موزیک مانند ملودی مشخص یا فراز و فرودهای یک قطعه کامل در آن چندان شکل نمی‌گرفت. با این حال، اهمیت Riffusion در نشان دادن یک مسیر جدید بود؛ یعنی بهره‌گیری از دستاوردهای مدل‌های تولید تصویر برای تولید موسیقی. این کار عملاً جرقه‌ای بود که پس از آن مدل‌های انتشار در موسیقی به شدت مورد توجه قرار گرفتند.

۴-۳-۲- MusicLM (مدل قدرتمند گوگل برای متن به موسیقی):

در اوایل ۲۰۲۳، محققان گوگل از مدل MusicLM رونمایی کردند که گامی بلند در کیفیت و توانایی تولید موسیقی از متن به‌شمار می‌رود. MusicLM یک مدل سلسله‌مراتبی مبتنی بر Transformer است که ایده‌های مدل موفق صوتی قبلی گوگل یعنی AudioLM را با شرط متن ترکیب کرده است. به طور خلاصه، MusicLM شامل سه سطح مدل خودبازگشتی است:

- مدل معنایی: که توالی توکن‌های معنایی موسیقی را بر اساس ویژگی‌های معنایی متن تولید می‌کند؛
- مدل آکوستیک خشن (coarse): که ویژگی‌های صوتی کلی (مانند یکتاهای کدک صوتی در سطح پایین‌تر) را از روی توکن‌های معنایی تولید می‌کند؛
- مدل آکوستیک ظریف (fine): که جزئیات نهایی موج صوتی را تکمیل می‌کند.

هر یک از این مدل‌ها یک Transformer Decoder بزرگ (حدود ۳۳۰ میلیون پارامتر در هر سطح) است که به ترتیب خروجی سطح قبل را مشروط به متن پیش‌بینی می‌کند. برای اینکه MusicLM بتواند از متن بهره ببرد، گوگل یک مدل کمکی به نام MuLan را به کار گرفت. MuLan یک مدل تعبیه مشترک متن–موسیقی است که توسط Huang و همکاران در ۲۰۲۲ ارائه شد [۸]. این مدل یک بردار ۱۲۸–بعدی مشترک برای یک قطعه موسیقی و توضیح متنی متناظر آن می‌آموزد، طوری که زوج متن–موسیقی واقعی به بردارهای نزدیک به هم نگاشت شوند. در MusicLM، از این خاصیت استفاده شده است: هنگام آموزش مدل، به‌جای متن، از تعبیه‌ی موسیقی MuLan (که متناظر موسیقی آموزش است) به‌عنوان شرط استفاده می‌شود و مدل یاد می‌گیرد موسیقی تولیدی‌اش با آن تعبیه (و لذا با متن اصلی) همخوان باشد. به بیان دیگر، MusicLM مشکل کمبود داده‌های زوج متن–موسیقی را با تکیه بر همین فضای مشترک برطرف کرد؛ چون برای آموزش نیاز نیست حتماً هر قطعه داده، کپشن متنی داشته باشد، بلکه هر قطعه موسیقی را می‌توان در فضای MuLan به بردار متناظرش نگاشت و از آن استفاده مشروط کرد. MusicLM با یک مجموعه عظیم شامل ۲۸۰ هزار ساعت موسیقی (حدود ۵ میلیون کلیپ صوتی) آموزش یافته است. خروجی مدل نرخ نمونه‌برداری ۲۴ kHz دارد و می‌تواند تا حدود چند دقیقه موسیقی پیوسته تولید کند. طبق ارزیابی‌های انجام‌شده، MusicLM کیفیت صوتی بسیار بالا و انسجام بلندمدت قابل قبولی ارائه می‌کند و در آزمون‌های مقایسه‌ای توسط شنوندگان، در هر دو بعد کیفیت و تطابق با متن بهتر از مدل‌های قبلی (مانند Riffusion و یک مدل به‌نام Mubert) عمل کرده است [۱۶]. به عنوان مثال، در یک آزمایش ترجیح‌دهی، موسیقی‌های تولیدی MusicLM در ۵۸٪ موارد نسبت به Riffusion ترجیح داده شدند که اختلاف معنی‌داری بود [۱]. همچنین توانایی ویژه‌ای که MusicLM نمایش داد، پردازش همزمان ملودی و متن بود: یعنی می‌توان به مدل یک ملودی زمزمه‌شده یا صوتی داد و از سوی دیگر یک توضیح متنی سبک/ژانر، و مدل ملودی ورودی را در سبک موردنظر "بازآفرینی" می‌کند (مثلاً آوازی را که کاربر با صدای خودش زمزمه کرده به یک قطعه ارکسترال حماسی تبدیل می‌کند). این قابلیت حاصل معماری چندمرحله‌ای هوشمند MusicLM است که در سطح معنایی امکان پیوند ملودی و متن را فراهم کرده است. یکی از نگرانی‌هایی که پیرامون MusicLM مطرح شد، موضوع حق کپی و شباهت به آثار موجود بود. مدلی با این حجم عظیم داده احتمال داشت بخش‌هایی از موسیقی آموزش‌دیده را حفظ کند و عیناً بازتولید نماید. پژوهشگران گوگل برای بررسی این موضوع، تحلیلی انجام دادند و خوشبختانه نشان دادند احتمال حفظ‌کردن طولانی‌مدت قطعات خاص در MusicLM بسیار پایین است. با این حال، برای اطمینان و نیز احتمال مشکلات حقوقی، گوگل مدل MusicLM را به طور عمومی منتشر نکرد (تنها مجموعه‌داده MusicCaps که شامل ۵۵۰۰ زوج کلیپ موسیقی و شرح متن است را منتشر کرد تا پژوهشگران دیگر بتوانند مدل‌های متن–به–موسیقی را ارزیابی کنند).

۵-۳-۲- مدل‌های متن–به–موسیقی متا (AudioCraft: MusicGen):

شرکت متا (فیس‌بوک) نیز در سال ۲۰۲۳ با پروژه‌ای به نام AudioCraft وارد این عرصه شد که شامل چند مدل متن–به–صوت مختلف بود. در این میان مدل MusicGen به طور خاص برای تولید موسیقی آموزش یافته است. MusicGen از برخی جهات شبیه MusicLM است اما با رویکرد ساده‌تر: MusicGen یک معماری تک‌مرحله‌ای دارد، بدین معنی که به جای سلسله مدل‌های جداگانه، یک ترنسفورمر خودبازگشتی واحد تمامی کدهای صوتی را به صورت همزمان تولید می‌کند [۹]. برای این منظور، MusicGen از یک استراتژی نوآورانه به‌نام نحوه‌ی درهم‌آمیزی توکن‌ها (efficient token interleaving) استفاده می‌کند که به مدل اجازه می‌دهد توکن‌های چندین کدبوک صوتی را به شکل یک دنباله‌ی واحد ولی ساخت‌یافته تولید کند. بنابراین برخلاف MusicLM که مثلاً ابتدا توکن‌های معنایی بعد آکوستیک خشن و ... را جداگانه تولید می‌کرد، MusicGen عملاً همه را در یک پاس پیش‌بینی می‌کند و نیاز به مدل‌های چندمرحله‌ای یا upsampling جداگانه ندارد. این باعث سادگی معماری و کاهش سربار محاسباتی می‌شود. MusicGen نیز همانند AudioGen، از کدک EnCodec برای برداری‌سازی صوت و یک انکدر متنی (مدل ترنسفورمری که روی مجموعه متن–آهنگ‌هایی آموزش دیده) برای فهم پرامپت بهره می‌گیرد. اندازه مدل‌های MusicGen در چند نسخه (کوچک تا بزرگ) ارائه شد که بزرگ‌ترینش حدود ۱.۵ میلیارد پارامتر داشت. نکته قابل توجه در MusicGen انتشار کد و مدل‌ها به‌صورت متن‌باز بود. متا وزن‌های مدل MusicGen را منتشر کرد و نشان داد که حتی با داده‌های کمتر (نسبت به MusicLM)، می‌توان نتایج قابل قبولی گرفت. مدل MusicGen بر روی ۴۰۰ ساعت موسیقی تحت لیسانس (شامل سبک‌های متنوع) آموزش داده شده و می‌تواند کلیپ‌هایی تا حدود ۳۰ ثانیه تولید کند. در ارزیابی‌های تطبیقی، MusicGen کیفیت صوتی نزدیک به MusicLM داشته و از آنجا که به صورت آزاد در دسترس است، به سرعت مورد توجه جامعه پژوهشی قرار گرفت. MusicGen همچنین اجازه می‌دهد یک ملودی راهنما نیز به همراه متن به آن داده شود (مشابه قابلیت MusicLM). در واقع دو نسخه خاص از این مدل با نام‌های MusicGen-Melody و MusicGen-Style توسط متا ارائه شد که یکی علاوه بر متن، یک فایل صوتی (ملودی اولیه) را به عنوان شرط ورودی می‌گیرد و دیگری قابلیت اعمال یک آهنگ مرجع برای تقلید سبک را دارد. این قابلیت‌ها انعطاف مدل را در کاربردهای مختلف نشان می‌دهد.

۶-۳-۲- مدل‌های دیفیوژن پیشرفته یعنی Noise2Music، Moûsai و JEN-1:

پس از موفقیت Riffusion و AudioLDM، پژوهشگران شروع به بهبود مدل‌های انتشار ویژه‌ی موسیقی کردند. Huang و همکاران (۲۰۲۳) مدل Noise2Music را معرفی کردند که ساختاری دومرحله‌ای داشت [۱۰]. مرحله اول یک دیفیوژن مولد برای ایجاد یک نمایش میانی از موسیقی بود و مرحله دوم یک دیفیوژن تکمیلی (Cascade) برای بهبود کیفیت و رزولوشن خروجی. آن‌ها دو نوع نمایش میانی را آزمودند: یکی اسپکترگرام و دیگری موج صوتی با نرخ نمونه‌برداری پایین (۳.۲ kHz). یافته جالب این بود که وقتی نمایش میانی را موج خام (ولی کم‌کیفیت) درنظر گرفتند، نتیجه‌ی نهایی بهتر از حالتی بود که نمایش میانی اسپکترگرام بود. Noise2Music توانست قطعات نسبتاً بلند (۳۰ ثانیه) با کیفیت ۲۴ kHz تولید کند. مدل دیگری به نام Moûsai توسط Schneider و همکاران (۲۰۲۳) ارائه شد که آن هم دومرحله‌ای بود اما با رویکرد متفاوت. Moûsai ابتدا سیگنال صوتی را توسط یک autoencoder انتشارمحور فشرده می‌کند تا بردار نهفته‌ی فشرده به دست آید؛ سپس در مرحله دوم یک دیفیوژن در فضای نهفته مشروط به متن اجرا می‌شود تا بردار نهفته‌ی خروجی را تولید کند. Moûsai موفق شد موسیقی استریو ۴۸ kHz به طول چند دقیقه تولید کند که از نظر کیفیت و ساختار بلندمدت بسیار چشمگیر بود [۱۱]. همچنین گزارش شده که این مدل بهینه‌سازی‌های زیادی در کد داشته و می‌تواند به صورت بلادرنگ (real-time) روی یک کارت گرافیک مصرفی نمونه‌سازی کند، که نشان‌دهنده‌ی پیشرفت در کارایی مدل‌های انتشار است. در ۲۰۲۴ نیز مدل JEN-1 معرفی شد. JEN-1 یک مدل انتشار همه‌منظوره موسیقی بود که قابلیت‌های مولتی‌تسک داشت؛ یعنی هم تولید موسیقی از متن انجام می‌داد، هم می‌توانست وظایفی مثل ترمیم موسیقی (inpainting) یا ادامه‌دادن یک قطعه ناقص (continuation) را انجام دهد. کلید موفقیت JEN-1 در یک معماری ترکیبی خودبازگشتی/غیرخودبازگشتی بود. به این صورت که بخش خودبازگشتی آن وابستگی‌های زمانی بلندمدت موسیقی را یادمی‌گیرد (مثلاً تکرار الگوها و پیشروی ملودیک) و بخش غیرخودبازگشتی آن تولید موازی بخش‌هایی از توالی را امکان‌پذیر می‌کند تا سرعت تولید بالا رود. این ترکیب باعث شده JEN-1 بتواند بدون افت کیفیت، در زمان مناسبی موسیقی ۴۸ کیلوهرتز استریو تولید کند. به علاوه، JEN-1 مستقیماً روی موج خام کار می‌کند و از واسط طیف‌نگاشت عبور نمی‌کند که این هم به بهبود وفاداری صوت خروجی کمک کرده است.

۴-۲- جمع‌بندی مدل‌ها و مقایسه:

جدول (۱-۴-۲) منتخبی از مدل‌های مطرح متن-به-موسیقی را خلاصه و مقایسه می‌کند، همراه با سال ارائه، نوع مدل، نوع ورودی/خروجی و ویژگی شاخص هر کدام که از نظر رویکرد فنی و قابلیت‌ها تنوع زیادی دارند. برخی (مانند TransProse یا مدل‌های HMM) نمادین هستند و خروجی MIDI تولید می‌کنند، در حالی که بسیاری از مدل‌های جدید (MusicLM، Riffusion، MusicGen و ...) مستقیماً موج صوتی واقعی را می‌سازند. همچنین ورودی برخی مدل‌ها فقط متن ساده است، اما برخی دیگر ورودی‌های اضافه مانند ملودی اولیه (برای هدایت بیشتر) یا برچسب‌های احساسی/سبک را نیز می‌پذیرند. از نظر تکنیک، نسل جدید تقریباً همگی بر پایه شبکه‌های ترنسفورمر یا انتشار هستند و شبکه‌های RNN سنتی کمتر مورد استفاده‌اند. معیار ارزیابی خروجی مدل‌ها نیز عموماً شامل ارزیابی ذهنی توسط انسان (مقایسه ترجیح یا امتیازدهی به تطابق و کیفیت) و معیارهای عینی مانند فاصله فروشه صوتی (FAD)، یا نرخ‌های خطا در طبقه‌بندی ژانر و غیره است. به طور کلی، روند حرکت پژوهش‌ها به سوی مدل‌های بزرگ‌تر با داده‌های بیشتر و ادغام بهتر دانش زبانی و موسیقایی است.

مدل (سال)	نوع مدل و معماری	ورودی‌ها	خروجی و بازنمایی	ویژگی‌ها و قابلیت‌ها
TransProse (2014)	قواعد-محور (قوانین احساسی)	متن رمان (انگلیسی)	پیانو MIDI ساده	نگاشت احساسات متن به گام و تمپو؛ تولید ملودی ساده بر اساس چگالی کلمات احساسی
Monteith et al. (2012)	HMM آماری برای هر احساس	برچسب احساس متن (دسته)	نت‌های ملودی (تک‌صدایی)	یادگیری ملودی‌های کوتاه برای هر طبقه احساس؛ استفاده از n-gram برای ریتم
Ferreira & Whitehead (2019)	عمیق (mLSTM + Logistic)	MIDI بدون + برچسب **برچسب مثبت/منفی**	MIDI چندصدایی (پیانو)	کشف نرون احساس در LSTM و قابلیت تنظیم خروجی به مثبت/منفی؛ مجموعه‌داده VGMIDI.
Jukebox (OpenAI 2020)	عمیق (VQ-VAE + Transformer)	**متن اشعار** + سبک/ژانر	صوت خام ۴۴kHz ترانه با (آواز)	تولید موسیقی با خواننده بر اساس اشعار؛ مدل عظیم (۵ میلیارد پارامتر)؛ نیاز به محاسبات سنگین
DiffSound (2022)	عمیق (CLIP + Diffusion)	متن (CLIP متن)	Mel-Spectrogram صوت	مدل انتشار روی طیف‌نگاشت؛ تولید افکت‌ها و صداهای محیطی؛ محدود به کلیپ‌های کوتاه
AudioGen (2022/23)	عمیق (T5 + Transformer)	متن توصیفی (انگلیسی)	کدک EnCodec (24 kHz)	خودبازگشتی روی توکن‌های صوتی؛ پشتیبانی از چندمنبع صدا؛ **بهبود با guidance**
Riffusion (2022)	عمیق (Stable Diffusion)	متن یا شعر آهنگ	صوت ۴۴kHz (چند ثانیه)	انتشار روی تصویر طیف‌نگاشت؛ **بلادرنگ و تعاملی**؛ محدودیت در طول و ساختار موسیقی
MusicLM (2023)	عمیق (3×Transformer (سلسله‌مراتبی	متن + (امکان ملودی زمزمه)	صوت ۲۴ kHz (دقیقه‌ها)	سه مرحله‌ی معنایی-صوتی؛ استفاده از تعبیه MuLan با تطابق عالی و منتشر نشده
MusicGen (2023)	عمیق (Transformer (تک‌مرحله‌ای	متن + (امکان ملودی/صوت راهنما)	کدک EnCodec 32 kHz	مدل متن‌یاز متا؛ بدون نیاز به چند مدل مجزا (همه در یکی)؛ تولید سریع‌تر موازی.
Noise2Music (2023)	عمیق (Diffusion (دومرحله‌ای	(T5) متن	صوت ۲۴ kHz (۳۰ ثانیه)	موج انتشار در دو سطح و کیفیت بهتر با نمایش موج خام
Moûsai (2023)	عمیق (AutoEnc + Diffusion)	(T5) متن	صوت ۴۸ kHz استریو (چند دقیقه)	فشردسازی با انتشار انکودر و انتشار نهفته با شرط متن؛ **بلادرنگ روی گرافیک خانگی**
JEN-1 (2024)	عمیق (Diffusion ترکیبی)	متن + (امکان ورودی صوت)	صوت ۴۸ kHz استریو	معماری خودبازگشتی/غیرخودبازگشتی؛ چندنقشه (نسل، ترمیم، ادامه)؛ عدم نیاز به طیف‌نگاشت
...	...	...	...	...

جدول (۱-۴-۲) : رویکرد فنی مدل‌ها

۳- معیارهای ارزیابی مدل‌ها

ارزیابی کیفیت و تطابق موسیقی تولیدشده با متن ورودی، یکی از چالش‌های اساسی در حوزه تبدیل متن به موسیقی (Text-to-Music)

است. برای این منظور، معیارهای کمی و کیفی متعددی پیشنهاد شده‌اند که در این بخش به مهم‌ترین آن‌ها می‌پردازیم. این معیارها شامل ارزیابی‌های کیفیت صوتی، تطابق محتوا با متن و ترجیح انسانی هستند. در ادامه هر معیار به‌همراه فرمول ریاضی (در صورت کاربرد) و نحوه محاسبه آن توضیح داده شده است.

۳-۱- MOS (Mean Opinion Score) – امتیاز میانگین نظرات

MOS یک معیار کیفی مبتنی بر نظرسنجی انسانی است که کیفیت ادراک‌شده‌ی موسیقی یا صدا را روی مقیاس مرتب (معمولاً ۱ تا ۵) اندازه‌گیری می‌کند [۱۸]. برای محاسبه‌ی MOS، تعدادی شنونده‌ی انسانی هر کلیپ صوتی را مثلاً بین ۱ (بسیار بد) تا ۵ (عالی) رتبه‌بندی می‌کنند. سپس میانگین حسابی این رتبه‌ها به‌عنوان MOS گزارش می‌شود. فرمول کلی MOS به صورت زیر است:

$$MOS = \frac{\sum_{n=1}^{N} R_n}{N}$$

که در آن Rn امتیاز شنونده nام و N تعداد کل شنوندگان است. MOS به دلیل ماهیت انسانی‌اش معیار نهایی ارزیابی کیفیت به‌شمار می‌رود اما انجام آن پرهزینه و زمان‌بر است. به همین دلیل، سایر معیارهای خودکار معرفی شده‌اند تا به‌عنوان جایگزین یا مکمل MOS به‌کار روند.

۳-۲- FAD (Fréchet Audio Distance) – فاصله‌ی صوتی فرشه

FAD یک معیار کمی مرجع–آزاد برای ارزیابی کیفیت کلی و واقع‌نمایی سیگنال‌های صوتی تولیدشده است [۱۵]. ایده‌ی FAD مقتبس از معیار Fréchet Inception Distance در تصاویر است که تفاوت توزیع آماری ویژگی‌های صوتی تولیدشده را با توزیع ویژگی‌های صوتی واقعی می‌سنجد. برای محاسبه‌ی FAD، ابتدا یک مجموعه مرجع از قطعات موسیقی واقعی و یک مجموعه ارزیابی از خروجی‌های مدل را در یک فضای ویژگی نهفته (مثلاً بردارهای ویژگی یک مدل مانند VGGish یا AudioCLAP) نمایش می‌دهیم [۱۹]. سپس با فرض توزیع گاوسی برای این ویژگی‌ها، فاصله فرشه بین دو توزیع گاوسی متناظر با مجموعه مرجع (میانگین μx، کواریانس Σx) و مجموعه تولیدی (میانگین μy، کواریانس Σy) محاسبه می‌شود. فرمول کلی فاصله فرشه بین دو توزیع گاوسی به صورت زیر است:

$$\mathrm{FAD}^2(X,Y) = \|\mu_X - \mu_Y\|_2^2 + \mathrm{tr}\left(\Sigma_X + \Sigma_Y - 2\sqrt{\Sigma_X \Sigma_Y}\right)$$

مقدار FAD پایین‌تر به معنی نزدیک‌تر بودن توزیع خروجی مدل به توزیع داده‌های واقعی و در نتیجه کیفیت بالاتر و واقع‌نمایی بیشتر است. به عنوان مثال، FAD نزدیک صفر نشان می‌دهد که صوت‌های تولیدشده از نظر آماری شبیه موسیقی‌های استودیویی بی‌نویز هستند. مزیت FAD این است که بدون نیاز به وجود مرجع مستقیم برای هر نمونه (مرجع آزاد) می‌تواند کیفیت کلی مجموعه‌ای از خروجی‌ها را بسنجد. با این حال، FAD محدودیت‌هایی دارد: فرض گاوسی بودن توزیع ویژگی‌ها در همه حالات برقرار نیست و برای مجموعه داده‌های کوچک اریب دارد و همچنین محاسبات آن با افزایش بعد فضای ویژگی و تعداد نمونه‌ها بسیار سنگین می‌شود [۱۵]. مطالعات نشان داده‌اند که در زمینه تولید موسیقی آزاد، همبستگی FAD با ترجیحات واقعی انسان پایین است. پس پژوهشگران به‌دنبال بهبود این معیار یا معرفی معیارهای جایگزین هستند [۲۰].

۳-۳- CLAP Score – امتیاز تطابق CLAP

مدل CLAP (Contrastive Language-Audio Pretraining) یک مدل تعبیه‌ساز متنی–صوتی است که بردارهای نهفته متن و صوت را در فضای مشترکی قرار می‌دهد. از این مدل می‌توان برای سنجش میزان تطابق موسیقی تولیدشده با متن توصیفی استفاده کرد [۱۹]. معیار امتیاز CLAP بدین صورت تعریف می‌شود که ابتدا بردار نهفته متن ورودی fT و بردار نهفته صوت تولیدشده fA را با مدل CLAP استخراج می‌کنیم. سپس شباهت کسینوسی میان این دو بردار را محاسبه کرده و میانگین می‌گیریم. فرمول این معیار به صورت زیر است:

$$\frac{{}_{\text{audio}}f_{\text{text}} \cdot f}{\|{}_{\text{audio}}f_{\text{text}}\| \, \|f\|} = \cos\left(f_{\text{text}}(T),\ f_{\text{audio}}(A)\right) = \text{CLAP-Score}(T, A)$$

امتیاز بالاتر CLAP نشان‌دهنده‌ی انطباق بیشتر موسیقی تولیدی با محتوای متن است. این معیار در واقع مشابه سنجش تطابق متن و تصویر توسط CLIP در حوزه‌ی بینایی است. به عنوان مثال، اگر متن درباره «ملودی شاد با گیتار آکوستیک» باشد، موسیقی تولیدی که امتیاز CLAP بالاتری کسب کند احتمالاً حاوی الگوهای صوتی مرتبط با ملودی شاد و صدای گیتار است. توجه داریم که CLAP Score یک معیار محتوایی است و کیفیت فنی صوت (مانند وضوح) را مستقیماً نمی‌سنجد بلکه بیشتر بر همخوانی محتوا با متن تاکید دارد. در پژوهش‌ها معمولاً CLAP Score در کنار FAD گزارش می‌شود تا توازن کیفیت صوتی و تطابق محتوایی بررسی گردد.

علاوه بر MOS، FAD و CLAP، معیارهای دیگری نیز برای ارزیابی مدل‌های تولید موسیقی استفاده می‌شوند و هر یک از این معیارها بخشی از عملکرد مدل را می‌سنجد و هیچ‌کدام به تنهایی تصویر کاملی ارائه نمی‌دهند. از این رو در کارهای پژوهشی اخیر معمولاً ترکیبی از معیارهای مذکور گزارش می‌شود تا جنبه‌های مختلف کیفیت و تطابق در نظر گرفته شود:

۳-۴- دقت بازیابی یا R-Precision:

که میزان توانایی مدل در تولید موسیقی منطبق با یک توصیف خاص را به‌صورت بازیابی صحیح در میان چندین گزینه می‌سنجد. برای محاسبه‌ی آن، معمولاً برای هر قطعه‌ی موسیقی تولیدی چند توضیح متن وجود دارد و باید بررسی شود توضیح درست در بین نزدیک‌ترین تعبیه‌های متنی به تعبیه‌ی صوتی قرار می‌گیرد یا خیر. هرچه مدل تطابق بهتری ایجاد کند، R-Precision بالاتر خواهد بود.

۳-۵- تنوع و پوشش (Diversity):

معیارهایی مثل درصد نغمه‌ها یا سازهای یکتا در خروجی‌ها یا فاصله‌ی پوشش توزیع ویژگی‌ها، برای سنجش متنوع بودن خروجی‌های مدل به‌کار می‌روند [۲۰]. یک مدل خوب نباید همه ورودی‌ها را به خروجی‌های بسیار مشابه تبدیل کند؛ بلکه باید انعطاف داشته باشد و طیفی متنوع از حالات موسیقایی را پوشش دهد. به عنوان مثال، تولید موسیقی‌های متفاوت در ژانرها و سرعت‌های گوناگون نشان‌دهنده‌ی تنوع بالاتر مدل است.

۳-۶- طول زمینه (Context Length):

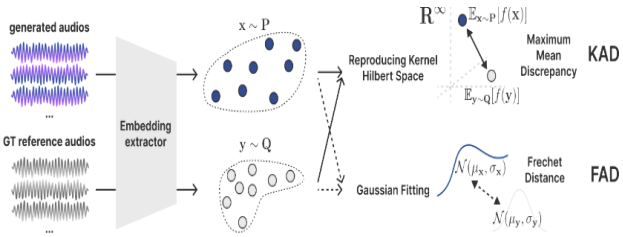
توانایی مدل در حفظ تداوم موسیقایی در بازه‌های طولانی نیز اهمیت دارد [۲۰]. برخی پژوهش‌ها با تخریب کنترل‌شده‌ی سیگنال (مثلاً کوتاه کردن زمینه‌ی قبل) و سنجش تغییر رتبه‌بندی ترجیح، حساسیت مدل به طول زمینه را ارزیابی می‌کنند.

۳-۷- ارزیابی کیفی با روش‌های مقایسه‌ای:

مانند انجام آزمون ترجیح دوتایی (A/B test) بین خروجی دو مدل مختلف توسط انسان‌ها، یا ارزیابی MUSHRA (در حوزه صدا برای مقایسه چند نمونه با مرجع) که جزو شیوه‌های کیفی مرسوم در سنجش موسیقی تولیدی هستند. این روش‌ها مکمل MOS بوده و به‌ویژه برای مقایسه‌ی مستقیم دو سیستم به‌کار می‌روند.

۳-۸- KAD (Kernel Audio Distance) – فاصله‌ی کرنل صوتی (معیار جدید)

یکی از جدیدترین معیارهای پیشنهادشده برای ارزیابی کیفیت صوتی مدل‌های مولد، Kernel Audio Distance (KAD) است [۱۵]. این معیار با شعار "No More FAD!" معرفی شده و تلاش دارد محدودیت‌های FAD را برطرف کند.



شکل (۸-۳) : مقایسه بین KAD (فاصله صوتی هسته) و FAD (فاصله صوتی فرشه).

مطابق شکل (۸-۳) KAD یک معیار مستقل از توزیع است که برای جاسازی توزیع‌های P و Q به هیچ فرض اساسی نیاز ندارد و بر پایه‌ی فاصله‌ی حداکثر میانگین (MMD) در فضای ویژگی صوت تعریف می‌شود. به عبارت دیگر، به جای فرض گاوسی بودن توزیع ویژگی‌ها، از روش غیرپارامتری MMD برای سنجش اختلاف دو توزیع (مرجع و تولیدی) استفاده می‌کند که هیچ فرض خاصی روی شکل توزیع ندارد. فرمول کلی MMD بین دو توزیع P و Q به صورت زیر است:

$$\mathrm{MMD}^2(P,Q) = \mathbf{E}_{x,x'}[k(x,x')] + \mathbf{E}_{y,y'}[k(y,y')] - 2\,\mathbf{E}_{x,y}[k(x,y)]$$

k(...) یک تابع کرنل مثبت تعریف‌شده (مثلاً کرنل گوسی RBF) است. این کمیت در صورتی‌که کرنل characteristic انتخاب شود، فاصله‌ای است که فقط وقتی دو توزیع یکسان باشند مقدار صفر می‌گیرد و در غیر این صورت مقداری مثبت متناسب با اختلاف توزیع‌ها خواهد داشت. KAD در واقع پیاده‌سازی همین MMD روی بردارهای تعبیه‌ی صوتی است. برای محاسبه‌ی KAD، مجموعه‌ی ویژگی‌های صوتی مرجع و مجموعه‌ی ویژگی‌های صوتی تولیدی را در یک فضای تعبیه (مانند CLAP یا AudioMAE) درنظر گرفته و با یک کرنل مثلاً RBF، مقدار MMD بین دو مجموعه محاسبه می‌شود. به‌طور پیش‌فرض کرنل RBF با عرض باند انتخاب‌شده بر اساس فاصله میانه بین نمونه‌های مرجع استفاده شده است که نیاز به تنظیم دستی را کاهش می‌دهد. مزایای گزارش‌شده برای KAD به‌طور خلاصه عبارتند از: توزیع–غیروابسته بودن (عدم نیاز به فرض نرمال)، ناریب بودن حتی در نمونه‌های کم (نیاز ندارد مانند FAD برای نمونه‌های کوچک اصلاح بایاس انجام شود) و کارایی محاسباتی بالاتر به ویژه در ابعاد بالای ویژگی. پیچیدگی محاسباتی KAD از مرتبه توان دوم n (به دلیل محاسبه کرنل بین جفت‌نمونه‌ها) است ولی قابل موازی‌سازی روی GPU بوده و عملاً برای مجموعه‌های معمول داده سریع‌تر از FAD گزارش شده است. مهم‌تر اینکه مطالعات نشان داده‌اند همبستگی KAD با قضاوت ادراکی انسان نسبت به FAD قوی‌تر است. بطور مشخص، در یک آزمون همبستگی اسپیرمن با رتبه‌بندی کیفی انسان‌ها، KAD امتیاز بالاتری نسبت به FAD کسب کرده است (نمودارهای مربوطه نشان می‌دهد KAD با حاشیه قابل توجهی به ترجیحات انسانی نزدیک‌تر است) [۱۵]. معرفی معیارهایی مانند KAD و MAD (Mauve Audio Divergence) نشان می‌دهد جامعه تحقیقاتی به کاوش در فضای معیارهای چندبعدی ادامه می‌دهد تا ارزیابی مدل‌های تولید موسیقی هرچه بیشتر با ترجیح شنوندگان انسانی همراستا شود.

۴- چالش‌ها و محدودیت‌ها

با وجود پیشرفت چشمگیر مدل‌های متن–به–موسیقی، این حوزه همچنان با چالش‌ها و موانع متعددی روبروست که در ادامه به برخی از مهم‌ترین آن‌ها اشاره می‌کنیم:

کمبود داده‌های موازی متن–موسیقی: مدل‌های یادگیری عمیق معمولاً برای عملکرد خوب نیاز به صدها هزار نمونه‌ی آموزشی دارند. در حوزه‌ی متن–به–موسیقی، جمع‌آوری مجموعه‌داده‌ای بزرگ که برای هر قطعه موسیقی توضیح متنی غنی داشته باشد، دشوار است. بیشتر موسیقی‌های در دسترس عملاً کپشن یا توصیف متنی دقیق ندارند. پژوهشگران برای رفع این مشکل به راه‌حل‌هایی مانند استفاده از فضای مشترک متن–موسیقی (مانند MuLan) یا برچسب‌گذاری خودکار و بهره‌گیری از داده‌های بدون برچسب روی آورده‌اند. با این حال، این چالش همچنان پابرجاست و حتی MusicLM با وجود circumvent کردن نیاز به متن در زمان آموزش، برای ارزیابی مجبور به ایجاد دیتاست MusicCaps شد.

کیفیت و تنوع داده‌های آموزشی: بسیاری از مدل‌های اخیر بر مجموعه‌های نسبتاً محدودی (از حیث سبک، سازبندی یا فرهنگ) آموزش دیده‌اند. برای مثال MusicGen بر ۴۰۰ ساعت موسیقی عمدتاً غربی آموزش دیده، لذا ممکن است در تولید سبک‌های موسیقی خارج از آن توزیع (مثل موسیقی سنتی شرق) دچار ضعف باشد. به علاوه، نگرانی‌های حق نشر باعث می‌شود نتوان از حجم عظیمی از موسیقی‌های تجاری در آموزش استفاده کرد. برخی پژوهش‌ها (مثل MusicLDM) به روش‌هایی مثل data augmentation (مانند Mixup همگام با ضرب‌آهنگ) متوسل شده‌اند تا ظرفیت خلاق مدل را بالاتر ببرند و آن را از حفظ‌کردن داده‌های محدود رها کنند. این روش ترکیب خطی تصادفی قطعات در زمان آموزش، به کاهش کپی‌برداری ناخواسته و افزایش تنوع کمک می‌کند.

همخوانی دقیق موسیقی با متن: اگرچه مدل‌های جدید در انتقال کلی حال‌وهوا و سبک متن موفق‌تر شده‌اند، اما هنوز جزئیات دقیق متن همیشه در موسیقی منعکس نمی‌شود. مثلاً MusicLM می‌تواند «ملودی ویولن آرام با بک‌گراند گیتار دیستورت» را تولید کند، اما اگر متن پیچیده‌تر باشد (مثلاً چند جمله شامل تغییرات پیاپی احساس)، مدل ممکن است فقط برداشت کلی را لحاظ کند و همه ظرائف را پوشش ندهد. این مسأله به ویژه وقتی متن داستان‌گونه یا چندوجهی باشد نمود می‌یابد. امکان کنترل موضعی (local) روی موسیقی متناسب با بخش‌های مختلف متن هنوز محدود است، هرچند برخی رویکردهای پژوهشی به سمتی می‌روند که مثلاً از مدل‌های alignment بین متن و زمان موسیقی استفاده کنند.

ساختار بلندمدت و انسجام زمانی: تولید قطعه موسیقی بلند (مثلاً چند دقیقه) که دارای ساختار آغاز، اوج و پایان و تکرار تم‌ها باشد، همچنان چالش‌برانگیز است. مدل‌های autoregressive مانند MusicLM از طریق طولانی کردن توالی تلاش کرده‌اند این مشکل را حل کنند اما حتی آن‌ها هم برای قطعات خیلی بلند ممکن است دچار سرگردانی ملودیک یا لوپ‌های تکراری ناخواسته شوند. مدل‌های انتشار نیز نیازمند حافظه طولانی هستند یا باید ترفندهای ویژه‌ای برای نگه‌داشتن تم اصلی در کل قطعه داشته باشند. رویکردهای ترکیبی (مانند JEN-1) سعی کرده‌اند با افزودن مولفه‌های خودبازگشتی، حافظه‌ی درازمدت را تقویت کنند، اما این حوزه جا برای پیشرفت دارد.

ارزیابی موسیقی تولیدی: سنجش کیفیت موسیقی به طور خودکار سخت است. برخلاف تصاویر که معیارهای نسبتاً خوبی (مثل FID) وجود دارد، در موسیقی معیارهای عینی کاملی در اختیار نیست. FAD (فاصله فروشه صوتی) یک معیار معمول است که توزیع آماری ویژگی‌های صوت تولیدی را با صوت واقعی مقایسه می‌کند، اما FAD پایین همیشه به معنای تطابق احساسی یا خلاقیت موسیقی نیست. در نهایت ارزیابی انسانی (میزان لذت‌بخش بودن موسیقی، میزان انطباق با متن) مهم‌ترین معیار باقی می‌ماند، که آن هم پرهزینه و زمان‌بر است. لذا پژوهشگران در تلاش‌اند متریک‌های بهتری ابداع کنند یا مدل‌های ارزیاب (مثلاً یک شبکه که ورودی‌اش متن و موسیقی است و خروجی‌اش امتیاز همخوانی) آموزش دهند؛ هرچند این کار هم چالش‌های خود را دارد. درنهایت KAD معیاری جدید بود که بهینه‌تر از روش‌های قبلی بود [۱۵].

کنترل‌پذیری و تفسیرپذیری: کاربران نهایی ممکن است بخواهند جزئیات بیشتری را کنترل کنند؛ مثلاً بگویند "همان ملودی را با تمپو کمی بیشتر و کمی غمگین‌تر تکرار کن". در حال حاضر، مدل‌ها بیشتر کنترل‌های سطح‌بالا دارند (توصیف کلی متن یا یک ملودی راهنما). افزودن واسطهای کنترلی تعاملی (مثل امکان ویرایش نت‌ها، یا تغییر احساس پس از تولید اولیه) یک زمینه پژوهشی مهم است. همچنین شفافیت مدل مطرح است: دانستن اینکه مدل چگونه از متن به موسیقی رسیده، هنوز شبیه به یک جعبه‌سیاه است. برخی کارهای اخیر (مثل AudioGenX) به سمت توضیح‌پذیر کردن شبکه‌های مولد صوتی رفته‌اند تا مشخص کنند کدام کلمات متن بر کدام بخش‌های موسیقی تاثیر گذاشته‌اند.

چالش‌های فنی (مقیاس و سرعت): بسیاری از مدل‌های مطرح بسیار بزرگ هستند (صدها میلیون تا چند میلیارد پارامتر) و آموزش آن‌ها نیازمند منابع محاسباتی عظیم است (مثلاً آموزش MusicLM به صدها هزار ساعت محاسبه GPU نیاز داشته است). تولید نمونه نیز خصوصاً برای مدل‌های autoregressive بلندمدت یا diffusion با مراحل زیاد، زمان‌بر است. تلاش‌هایی در جهت بهینه‌تر کردن وجود دارد، از جمله استفاده از مدل‌های فشرده‌تر، روش‌های افزایش سرعت نمونه‌گیری (sampling) و بهره‌گیری از واحدهای محاسباتی خاص برای صوت. برای مثال، مدل Moûsai تاکید خاصی بر بهینه‌سازی داشت و توانست نمونه‌سازی را سریع کند. با این حال، قابل‌دسترس کردن این فناوری برای عموم (روی یک لپ‌تاپ معمولی) هنوز کاملاً محقق نشده است.

با در نظر گرفتن موارد فوق، روشن است که اگرچه مسیر زیادی طی شده، اما موانع علمی و فنی قابل توجهی نیز پیش روست. در بخش نتیجه‌گیری به برخی راهکارهای ممکن و جهت‌های آینده اشاره می‌کنیم.

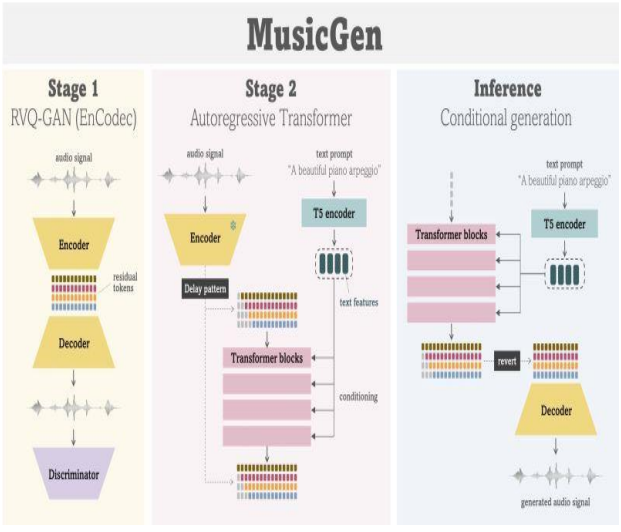
۵- نمونه پیاده‌سازی (تبدیل متن به موسیقی)

برای درک بهتر نحوه کار یک مدل متن–به–موسیقی، در این بخش یک پیاده‌سازی ساده با استفاده از کتابخانه‌های متن‌باز ارائه می‌شود. یکی از مدل‌های در دسترس، مدل MusicGen متا است که وزن‌های آموزش‌دیده‌ی آن توسط HuggingFace منتشر شده است [۲۱]. کد زیر به زبان پایتون نشان می‌دهد چگونه می‌توان از این مدل برای تولید یک قطعه موسیقی استفاده کرد [۱۷]:

```
# نصب کتابخانه‌های لازم
!pip install transformers datasets scipy

from transformers import AutoProcessor, MusicgenForConditionalGeneration
from IPython.display import Audio

# بارگذاری مدل و پردازشگر MusicGen (نسخه کوچک)
processor = AutoProcessor.from_pretrained("facebook/musicgen-small")
model = MusicgenForConditionalGeneration.from_pretrained("facebook/musicgen-small")

# تعریف پرامپت متنی (می‌تواند فارسی یا انگلیسی باشد)
text_prompt = "A calm Persian classical music with santur and a slow rhythm"

# آماده‌سازی ورودی مدل
inputs = processor(text=[text_prompt], return_tensors="pt")

# تولید توکن‌های صوتی به صورت خودبازگشتی
audio_tokens = model.generate(**inputs, do_sample=True, guidance_scale=3.0, max_new_tokens=256)

# به موج صوتی و پخش آن (tokens) تبدیل خروجی مدل
audio_array = audio_tokens[0].numpy()  # استخراج آرایه‌ی صوتی از خروجی مدل
sampling_rate = model.config.audio_encoder.sampling_rate  # نرخ نمونه‌برداری (مثلاً ۳۲۰۰۰ هرتز)
Audio(audio_array, rate=sampling_rate)
```

کد (۱-۴) : پیاده‌سازی نمونه

در کد (۱-۴) که البته نسخه‌ی کامل آن در فایل نوت‌بوک ژوپیتر برای اجرا در کولب قرار دارد، ابتدا مدل و پردازشگر متن مربوط به MusicGen بارگذاری می‌شوند. سپس یک رشته متنی به عنوان پرامپت تعریف شده (در اینجا توضیحی به زبان انگلیسی درباره موسیقی کلاسیک ایرانی آرام با سنتور) و به مدل داده می‌شود. مدل MusicGen خروجی خود را به شکل یک بردار سه‌بعدی (batch, channels, length) از نمونه‌های صوتی (محدوده -۱ تا ۱) برمی‌گرداند. در اینجا با استفاده از IPython.display.Audio می‌توانیم مستقیماً صدای تولیدشده را بشنویم یا توسط تابع write از کتابخانه scipy آن را در فایل wav ذخیره کنیم. شایان ذکر است که مدل‌های دیگر نظیر Riffusion نیز با کتابخانه Diffusers قابل استفاده‌اند، اما برای سادگی اینجا MusicGen را انتخاب کردیم. به طور مشابه می‌توان به جای عبارت انگلیسی، یک متن فارسی نیز وارد کرد؛ هرچند مدل‌های فعلی عمدتاً روی متن انگلیسی آموزش دیده‌اند و درک محدودی از زبان فارسی دارند. برای مثال، اگر متن فارسی "یک قطعه موسیقی شاد با سازهای ایرانی و ریتم تند" را به مدل بدهیم، ممکن است نتیجه نزدیک به انتظار باشد اما تضمین‌شده نیست. این نمونه کد صرفاً جهت آشنایی است و اجرای موفق آن به داشتن سخت‌افزار مناسب (خصوصاً GPU) نیاز دارد. با این وجود، نشان می‌دهد که چگونه در چند خط می‌توان از قدرت یک مدل متن-به-موسیقی استفاده کرد و موسیقی تولید نمود. توسعه‌دهندگان می‌توانند با تنظیم پارامترهایی مثل guidance_scale (برای کنترل میزان انطباق با متن) و یا با دادن ملودی اولیه (در صورت استفاده از نسخه melody مدل MusicGen)، خروجی‌ها را تا حدی به سلیقه خود نزدیک‌تر کنند. در ادامه مطابق شکل(۱-۵) جزئیات معماری، مبانی ریاضی و نحوه‌ی عملکرد این مدل به طور کامل تشریح می‌شود.



شکل (۸-۳) : معماری مدل انتخابی MusicGen

۱-۵- برداری‌سازی صوت با EnCodec

برای اینکه یک ترنسفورمر بتواند موسیقی را تولید کند، ابتدا لازم است سیگنال صوتی به توکن‌های قابل‌فهم برای مدل تبدیل شود. MusicGen برای این منظور از یک کدک عصبی صوت به نام EnCodec استفاده می‌کند. EnCodec یک مدل Encoder-Decoder مبتنی بر شبکه‌های عصبی کانولوشنی است که صوت خام (برای مثال با نرخ نمونه‌برداری ۳۲ کیلوهرتز) را به یک بردار ویژگی فشرده در هر بازه‌ی زمانی تبدیل می‌کند. سپس با به‌کارگیری تکنیک کمّی‌سازی برداریResidual یا به اختصار RVQ، این بردارهای پیوسته به مقادیر گسسته نگاشت می‌شوند [۲۲]. کمّی‌سازی RVQ بدین صورت است که به جای تنها یک کدبوک بزرگ، چندین کدبوک پیاپی (Stacked Codebooks) برای بازنمایی بردار استفاده می‌شود: ابتدا اولین کدبوک نزدیک‌ترین بردار کد را انتخاب کرده و بازسازی می‌کند؛ سپس خطای باقیمانده (رزیدیو) توسط کدبوک دوم کمّی‌سازی می‌شود و الی آخر. خروجی EnCodec در واقع چند توکن گسسته موازی است که هر کدام از یک کدبوک مجزا آمده‌اند. به بیان دیگر، هر فریم زمانی صوت توسط مجموعه‌ای از شاخص‌های گسسته (یکی از هر کدبوک) نمایش می‌یابد. به عنوان مثال، EnCodec به‌کاررفته در MusicGen صوت تک‌کاناله‌ی kHz۳۲ را با یک مدل ۵لایه‌ی کانولوشنی (استریمی) به نرخ حدود ۵۰ فریم‌برثانیه می‌رساند و سپس بردار هر فریم را با ۴ کدبوک مجزا کمّی‌سازی می‌کند. بنابراین هر فریم (مثلاً هر ۲۰ میلی‌ثانیه) به ۴ عدد توکن گسسته تبدیل می‌شود و در مجموع هر ثانیه موسیقی توسط تقریباً ۵۰×۴ = ۲۰۰ توکن کدبکی نمایش داده می‌شود. این بازنمایی فشرده‌ی صوتی پایه‌ی مدل‌سازی MusicGen است؛ یعنی وظیفه‌ی مدل زبانی ترنسفورمر، تولید پیاپی این توکن‌های صوتی است تا در نهایت توسط دیکدر EnCodec به موج صوتی تبدیل شوند.

۲-۵- مدل‌سازی خودبازگشتی توکن‌ها و الگوهای درهم‌آمیزی

وجود چندین توکن موازی در هر گام زمانی (یکی در هر کدبوک) یک چالش مهم ایجاد می‌کند: چگونه باید این دنباله‌ی چندبعدی از توکن‌ها را به توالی یک‌بعدی جهت مدل‌سازی خودبازگشتی تبدیل کرد [۹]؟

ساده‌ترین راه تخت‌سازی کامل (Flattening) است، بدین معنی که توکن‌های همه‌ی کدبوک‌ها را پشت سر هم در یک رشته‌ی طولانی قرار دهیم. در این حالت، مدل ترنسفورمر یک توالی خطی (تعداد فریم‌ها × تعداد کدبوک‌ها) را مدل می‌کند و از نظر تئوری می‌تواند توزیع مشترک تمامی توکن‌ها را به طور دقیق بیاموزد.

ولی این فرض در عمل کاملاً برقرار نیست؛ کدبوک‌های موازی یک فریم معمولاً هم‌بستگی دارند (مثلاً نمایه‌های طیفی و زمانی مرتبط صوت را تشکیل می‌دهند). در نتیجه این تجزیه‌ی غیردقیق می‌تواند به افت کیفیت منجر شود، خصوصاً اگر یکی از کدبوک‌ها خطایی داشته باشد مدل نمی‌تواند آن را با سایر کدبوک‌های همان فریم جبران کند. البته این ساده‌سازی، سرعت تولید را به‌شدت افزایش می‌دهد و پیچیدگی را کاهش می‌دهد، از این رو ارزش بررسی دارد. متا برای بررسی سیستماتیک این trade-off (دقت در برابر پیچیدگی)، رویکردی موسوم به الگوهای درهم‌آمیزی توکن‌ها معرفی کرد. ایده این است که بین دو حالت تخت‌سازی کامل و موازی کامل، می‌توان الگوهای واسطی را تعریف کرد که در آن برخی کدبوک‌ها با هم و برخی با تاخیر زمانی پیش‌بینی شوند. . نتایج نشان داد که روش تخت‌سازی کامل بالاترین کیفیت را حاصل می‌کند اما به قیمت ۴ برابر مراحل بیشتر (و زمان محاسباتی طولانی‌تر). در مقابل، روش‌های موازی سرعت بالایی دارند ولی کمی افت کیفیت ایجاد می‌کنند. روش‌های میانی مانند «اول‌خشن» یا «تخت جزئی» (که در آن کدبوک اول به طور جداگانه یا نیمه‌جدا تولید

می‌شود) کیفیتی بینابین داشتند. نکته‌ی مهم این بود که یک الگوی ساده‌ی تأخیری می‌تواند تقریباً همان کیفیت تخت‌سازی را با هزینه‌ی محاسباتی بسیار کمتر فراهم کند. در الگوی تأخیری، کدبوک‌ها با تأخیر زمانی یک واحدی نسبت به یکدیگر پیش‌بینی می‌شوند. به عنوان مثال، برای هر فریم ابتدا توکن کدبوک ۱ در یک گام تولید می‌شود؛ سپس توکن کدبوک ۲ (همان فریم) در گام بعدی می‌آید، سپس کدبوک ۳ و ۴. بدین ترتیب توکن‌های یک فریم به جای استقلال کامل، با فاصله‌های کوتاه پشت سر هم مدل می‌شوند و وابستگی محدودی میانشان برقرار می‌گردد (توکن‌های کدبوک‌های قبلی یک فریم در حافظه مدل برای پیش‌بینی کدبوک‌های بعدی حضور دارند).

با وجود این کاهش چشمگیر، کیفیت خروجی تقریباً هم‌سطح باقی می‌ماند؛ چنان‌که پژوهش متا نشان داد کیفیت موسیقی حاصل از الگوی تأخیری تنها اختلاف جزئی با حالت تخت دارد و نسبت به سایر الگوریتم‌های موجود همچنان برتری دارد. بر همین اساس، MusicGen در پیاده‌سازی نهایی خود از الگوی تأخیری برای ترتیب‌دهی توکن‌های صوتی استفاده می‌کند(این همان «efficient token interleaving» است که در معرفی مدل به آن اشاره شده‌است).

۳-۵- معماری مدل ترنسفورمر MusicGen

پس از تعیین ترتیب توکن‌ها در توالی، یک مدل زبان خودبازگشتی (decoder) وظیفه‌ی یادگیری توزیع این توکن‌ها و تولید آن‌ها را بر عهده دارد [۹]. MusicGen از یک ترنسفورمر Decoder استاندارد با چندین لایه استفاده می‌کند که به شرط (پرامت) متنی یا ملودیک نیز حساس شده است. ورودی این ترنسفورمر توالی توکن‌های صوتی کمّی‌شده (تخت‌شده بر اساس الگوی انتخابی) است که ابتدا به بردارهای نهفته (embedding) تبدیل می‌شوند. اما بر خلاف یک مدل زبان معمولی، در اینجا هر گام ورودی می‌تواند شامل چندین توکن (از کدبوک‌های مختلف) باشد یا برعکس، برخی کدبوک‌ها در آن گام غایب باشند. MusicGen این مسئله را با یک سازوکار تعبیه‌سازی ویژه حل کرده است: برای هر کدبوک یک جدول تعبیه (embedding table) مجزا با ابعاد مشخص (مثلاً ۱۰۲۴) در نظر گرفته شده است. سپس تمامی بردارهای تعبیه‌ی حاصل (از هر کدبوک) با هم جمع برداری می‌شوند تا یک بردار نهفته‌ی یکتا برای ورودی آن مرحله ساخته شود. به این بردار، یک تعبیه‌ی مکانی (positional) نیز اضافه می‌شود که نشان‌دهنده‌ی موقعیت آن مرحله در توالی کل است (برای این منظور از موقعیت‌یابی سینوسی استاندارد استفاده شده است). نتیجه‌ی نهایی به عنوان ورودی لایه‌های ترنسفورمر به کار می‌رود. هر لایه‌ی ترنسفورمر MusicGen شامل دو بخش توجهی است: یک بلوک Self-Attention علی (causal) که وابستگی‌های زمانی توکن‌های گذشته را مدل می‌کند، و سپس یک بلوک Cross-Attention که سیگنال‌های شرطی را با توالی درونی ادغام می‌کند. منظور از سیگنال شرطی (conditioning) می‌تواند متن یا ملودی باشد که به مدل داده می‌شود(بدین معنی که مدل می‌تواند خروجی موسیقی را بر اساس ورودی‌های شرط متنی یا صوتی هدایت کند. شرط متنی همان توصیف زبان طبیعی موردنظر کاربر درباره‌ی موسیقی است. شرط صوتی نیز می‌تواند یک قطعه‌ی ملودی راهنما باشد تا مدل بر اساس آن ملودی ادامه‌ی موسیقی را بسازد یا سبک آن را تقلید کند مثل زمزمه یا فایل MIDI یا تک‌خط ملودی یک ساز). به‌عبارت دیگر، MusicGen یک معماری دیکدر-تنها (GPT-مانند) دارد که در هر لایه پس از انجام Self-Attention بر روی تاریخچه‌ی توکن‌های صوتی، بردارهای میانی را به کمک مکانیزم توجه-متقاطع به بردارهای شرط (متن یا ملودی) نیز وابسته می‌کند. خروجی هر لایه از طریق یک شبکه‌ی کاملاًمتصل (Linear-ReLU-Linear) پردازش شده و با رزیدوال اسکیپ‌کانکشن به ورودی آن لایه جمع می‌شود (ترنسفورمر با نورم پیش‌ازرزیدوال به‌کار رفته است). تعداد لایه‌های ترنسفورمر و اندازه‌های آن بسته به نسخه‌ی مدل متفاوت است (مثلاً MusicGen نسخه‌ی بزرگ حدود ۲۴ لایه و بعد مدل ۱۵۳۶ دارد). در نهایت، خروجی آخرین لایه‌ی ترنسفورمر در هر گام زمانی به چند سری لوجیت تبدیل می‌شود: برای هر کدبوکی که در آن گام حضور داشته باشد یک لایه‌ی خطی جداگانه وجود دارد که ابعاد بردار خروجی را به اندازه‌ی تعداد کدهای آن کدبوک نگاشت کرده و توزیع احتمالات نرم‌ماکس بر روی کدبوک مربوطه را به دست می‌دهد. به عنوان مثال، اگر در یک مرحله قرار است توکن کدبوک ۲ و ۳ تولید شوند، مدل دو بردار لوجیت همزمان می‌دهد: یکی به طول مثلاً ۱۰۲۴ برای انتخاب مقدار کدبوک ۲، و دیگری مشابه برای کدبوک ۳. در مرحله‌ی بعد (تولید خودبازگشتی)، از این توزیع‌ها یک نمونه (توکن) برداشته می‌شود و به دنباله‌ی تولیدشده ضمیمه می‌گردد و فرآیند برای گام بعدی تکرار می‌شود.

۴-۵- مقیاس مدل و داده‌ی آموزش

MusicGen در چند اندازه مختلف ارائه شده است. مدل کوچک (MusicGen-small) ۳۰۰ میلیون پارامتر دارد و جهت تست‌های سبک‌تر کاربرد دارد [۲۱]. مدل بزرگ استاندارد حدود ۱/۵ میلیارد پارامتر دارد که به عنوان مدل اصلی متن-به-موسیقی استفاده می‌شود. همچنین یک مدل بزرگ‌تر حدود ۳/۳ میلیارد پارامتری نیز آموزش داده شده که مختص ورودی متن (بدون ملودی) است. مدل Melody و Style ذکرشده نیز هر کدام ~۱/۵ میلیارد پارامتر دارند و بر پایه‌ی همان معماری ترنسفورمر با انکدر متنی T5 ساخته شده‌اند. تمامی این وزن‌ها و کد منبع مدل‌ها توسط متا به صورت متن‌باز منتشر شده است و در هاب HuggingFace قابل دسترس هستند.

داده‌ی آموزشی MusicGen شامل مجموعه‌ای محدود اما باکیفیت از موسیقی‌های دارای مجوز است. به گفته‌ی متا، این مدل تنها بر روی حدود چند صد ساعت موسیقی دارای لیسانس (شامل طیف گسترده‌ای از سبک‌ها) آموزش یافته است، در حالی که مدل‌هایی نظیر MusicLM از هزاران ساعت موسیقی (اغلب بدون مجوز مشخص) بهره گرفته بودند. با وجود داده‌ی کمتر، MusicGen به خوبی توانسته ساختارهای موسیقایی را یاد بگیرد و کیفیتی رقابتی ارائه دهد. در یک ارزیابی تطبیقی استاندارد (مجموعه MusicCaps گوگل)، MusicGen به امتیاز ۸۴/۸ از ۱۰۰ در ارزیابی کیفی انسانی دست یافت که بالاتر از مدل‌های پیشین مانند MusicLM (امتیاز ~۸۰/۵) بود. این برتری هم در وفاداری ملودیک و هم در تطابق با متن خود را نشان داد [۹]. از جنبه‌ی کیفیت صوتی خام، MusicGen با نرخ ۳۲ کیلوهرتز و تولید استریو توانست خروجی‌هایی با کیفیت نزدیک به صدای واقعی ارائه کند. متا حتی یک دیکدر بهبودیافته مبتنی بر diffusion برای EnCodec معرفی کرده که می‌تواند کیفیت صدا را باز هم بهتر کند (هرچند با هزینه‌ی محاسباتی بیشتر). به طور کلی MusicGen نشان داد که با یک مدل ترنسفورمر واحد می‌توان قطعات موسیقی چندثانیه‌ای (تا حدود ۳۰ ثانیه)

ساخت که برای گوش انسان خوشایند و معنادار هستند، آن‌هم در سبک‌های متنوع و با امکان کنترل توسط کاربر. انتشار آزادانه‌ی این مدل باعث شد به سرعت مورد توجه جامعه‌ی پژوهشی قرار گیرد و در پروژه‌های مختلف (از جمله تولید موسیقی پس‌زمینه، ابزارهای آهنگسازی کمکی و تحقیقات تطبیقی) به کار گرفته شود. مدل MusicGen در ابتدا برای صدای مونو آموزش دیده بود، اما پژوهشگران متا نشان دادند که چارچوب آن را می‌توان بدون افزایش پیچیدگی، به تولید استریو تعمیم داد. برای این منظور، همان EnCodec به‌صورت جداگانه روی کانال چپ و راست اعمال می‌شود و به جای ۴ کدبوک، در مجموع ۸ کدبوک (۴ کدبوک برای هر کانال) خروجی می‌دهد. سپس الگوی درهم‌آمیزی delay با دو گزینه به‌کار گرفته شد: یکی الگوی "تأخیری استریو" که بین کانال چپ و راست نیز عدم توازن زمانی می‌اندازد (مثلاً کدبوک۱ چپ کمی زودتر از کدبوک۱ راست تولید شود) و دیگری الگوی "تأخیری نیمه‌موازی" که در آن توکن‌های هر دو کانال برای کدبوک‌های رده پایین‌تر (مثلاً کدبوک‌های ۲ تا ۴) همزمان و با یک تأخیر نسبت به کدبوک۱ تولید می‌شوند.

طبق نتایج متا، خروجی استریوی MusicGen نسبت به خروجی مونو از نظر شنوندگان اندکی با کیفیت‌تر و واقعی‌تر ارزیابی شد، و بین دو الگوی فوق، الگوی تأخیری همزمان کمی برتری کیفی و تطابق متنی بیشتری نشان داد.

۶- انتقادات و بحث

وضعیت فعلی: مرور انجام‌شده نشان می‌دهد که حوزه‌ی تولید موسیقی بر اساس متن در پنج سال اخیر از مراحل ابتدایی به سمت بلوغ حرکت کرده است. مدل‌های اولیه که بر قواعد یا معماری‌های ساده مبتنی بودند، اکنون جای خود را به شبکه‌های عظیم ترنسفورمر و انتشار داده‌اند که قادر به تولید قطعات موسیقی با کیفیت قابل توجه هستند. به خصوص، مدل‌های ۲۰۲۲ به بعد (AudioGen، Riffusion، MusicLM، MusicGen و ...) موفق شده‌اند شکاف کیفیت بین موسیقی تولیدی ماشین و موسیقی واقعی را کمتر کنند. برای نخستین بار، سیستم‌هایی مانند MusicLM نشان دادند که می‌توان توصیف‌های پیچیده متنی را که شامل سازها، ژانر، حس و حتی ملودی دلخواه است، به موسیقی قابل قبول تبدیل کرد. همچنین در ارزیابی‌های مبتنی بر نظر انسان، این مدل‌ها اغلب بهبود معناداری نسبت به نسل‌های قبلی داشته‌اند – مثلاً MusicLM و AudioGen هر دو برتری‌هایی نسبت به مدل‌های انتشار پایه (مانند Riffusion) در آزمون‌های ترجیح شنوندگان کسب کردند. [۵] از منظر تحلیل احساسات، همپوشانی جالبی بین دستاوردهای NLP و موسیقی به وجود آمده است. مدل‌های زبانی بزرگ (LLM) اکنون بخشی از معماری برخی سیستم‌های متن–به–موسیقی شده‌اند (برای فهم بهتر متن و استخراج جزئیات از آن). به علاوه، پژوهش‌های ویژه در مورد کنترل احساسی موسیقی (مثل کار Ferreira 2019) نشان داده‌اند که می‌توان تا حدودی حالت عاطفی موسیقی را از طریق بردارهای نهان شبکه تنظیم کرد. با این حال، ادغام صریح تحلیل‌گر احساسات متن به عنوان یک ماژول مجزا در سیستم‌های اخیر کمتر دیده می‌شود؛ چرا که مدل‌های بزرگ به طور ضمنی بسیاری از این مفاهیم را یاد می‌گیرند. به عنوان نمونه، MusicLM نیازی نداشت که برچسب "شاد" یا "غمگین" به طور جداگانه به آن داده شود – اگر در متن کلماتی با بار احساسی مثبت باشد یا اصطلاحاً mood ذکر شود (مثل "calming" یا " distorted guitar riff" برای حس انرژی)، مدل خود آن را تفسیر و اعمال می‌کند. بنابراین، در مدل‌های End-to-End جدید، تحلیل احساسات به صورت خودکار در دل مدل نهفته است. با این وجود، در کاربردهای خاص همچنان ممکن است یک لایه‌ی کنترل احساسی صریح مفید باشد؛ برای مثال، کاربری که یک متن خنثی (مثلاً شعر) را وارد می‌کند شاید بخواهد انتخاب کند موسیقی نتیجه شاد باشد یا غمگین. اینجاست که ترکیب ماژول‌های جداگانه‌ی NLP (برای استنتاج احساس متن) با مدل‌های مولد می‌تواند کنترل‌پذیری بیشتری بدهد.

مقایسه رویکردها: روش‌های مبتنی بر ترنسفورمر خودبازگشتی (AudioGen، MusicGen، MusicLM) در تولید جزئیات زمانی موسیقی بسیار خوب عمل می‌کنند و ساختارهای پیچیده‌ای را می‌توانند یاد بگیرند. آن‌ها برای حفظ بلندی توالی از تکنیک‌هایی مثل چندجریانی یا سلسله‌مراتب استفاده کرده‌اند. [۵] روش‌های دیفیوژن (Riffusion، DiffSound، Moûsai) مزیت تولید موازی دارند و کیفیت بالای صوتی (وضوح طیفی) را به ارمغان آورده‌اند، اما نیازمند تکرارهای متعدد نمونه‌گیری هستند و کنترل آن‌ها اندکی غیرمستقیم‌تر است (مثلاً via guidance). به نظر می‌رسد ترکیب این دو رویکرد راهی امیدوارکننده باشد – همچنان که مدل MeLoDy تلاش کرد با هدایت یک دیفیوژن توسط مدل زبان، از مزایای هر دو بهره ببرد. همچنین رویکردهای جدیدتر بر مدل‌های تک‌مرحله‌ای تأکید دارند تا پیچیدگی اجرای چند مدل پشت‌سرهم را کاهش دهند. این به سادگی معماری و کاهش خطاهای تجمعی کمک می‌کند.

کیفیت خروجی: کیفیت موسیقی تولیدشده اکنون در برخی جنبه‌ها رضایت‌بخش است ولی هنوز تا سطح خلاقیت و ظرافت موسیقی انسان‌ساخت فاصله دارد. برای مثال، مدل‌ها می‌توانند سبک را تقلید کنند یا ترکیب‌کنند، ولی آیا می‌توانند نغمه‌ای کاملاً بدیع و گیرا بسازند؟ یا مثلاً سوپرایزهای هنری در قطعه ایجاد کنند؟ این موضوعات هنوز باز هستند. از نظر احساسات، مدل‌ها در بیان احساسات کلی (شاد، غمگین، حماسی، آرام و ...) موفق‌اند اما در انتقال احساسات پیچیده یا متناقض ممکن است ضعف داشته باشند. برای نمونه، انتقال حس "نوستالژی شیرین" یا "اضطراب همراه امید" احتمالاً فراتر از درک فعلی مدل‌هاست و نیاز به ظرافت در تغییرات مدالیتی موسیقی دارد.

۷- نتیجه گیری و چشم‌انداز آینده

در این مقاله مروری، تحول فناوری تولید موسیقی از متن با محوریت تحلیل احساسات مورد بررسی قرار گرفت. از سیستم‌های ابتدایی قواعدمحور که احساسات متن را با قواعد دستی به نُت‌ها نگاشت می‌کردند تا مدل‌های غول‌پیکر مبتنی بر شبکه‌های ترنسفورمر و انتشار که قادر به ساخت قطعات موسیقی غنی و چنددقیقه‌ای هستند، مسیر پژوهش به وضوح نشان‌دهنده‌ی پیشرفتی چشمگیر است. به طور خلاصه، نتایج به‌دست‌آمده را می‌توان چنین جمع‌بندی کرد:

ادغام موفق NLP و موسیقی: مدل‌های نسل جدید توانسته‌اند دانش زبانی (معانی کلمات، حالات احساسی، مفاهیم سبک‌شناختی) را با دانش موسیقایی تلفیق کنند و سیستم‌های واحدی بسازند که مستقیماً متن کاربر را به صدای موسیقی تبدیل می‌کنند. این پیشرفت حاصل دستاوردهای موازی در هر دو حوزه (مدل‌های زبانی بزرگ و مدل‌های مولد صوت) بوده است. به بیان

دیگر، بدون Transformers و توانایی آن‌ها در درک متن، و بدون VAEها و Diffusionها در تولید صوت، این حد از موفقیت میسر نبود.

بهبود چشمگیر کیفیت و تنوع: خروجی مدل‌های اخیر از نظر شنیداری به مراتب دلچسب‌تر و متنوع‌تر از نسل‌های قبل است. موسیقی تولیدشده نه تنها نویز کمتری دارد بلکه می‌تواند سازهای مختلف، ریتم و ملودی نسبتاً معنادار، و حتی آواز بشرگونه (در مورد مدل‌هایی مثل Jukebox) داشته باشد. هرچند هنوز هم تشخیص ماشینی بودن در بسیاری موارد ممکن است، اما فاصله رفته‌رفته کمتر می‌شود.

اهمیت داده و مقیاس: یکی از درس‌های کلیدی این حوزه (مانند بسیاری حوزه‌های دیگر AI) آن است که مقیاس مدل و داده نقش بسیار مهمی ایفا می‌کند. مدل‌هایی که با صدها هزار ساعت موسیقی آموزش دیده‌اند (MusicLM) توانایی‌هایی را بروز داده‌اند که در مدل‌های آموزش‌دیده با چند صد ساعت (مدل‌های کوچکتر) مشاهده نشد. بنابراین، یکی از راه‌های آینده احتمالاً رفتن به سمت مدل‌های حتی بزرگ‌تر (در حد GPT-4 یا بالاتر اما برای موسیقی) است. البته باید مکانیسم‌های Regularization و جلوگیری از حفظ عین داده نیز تقویت شود تا مسائل حق‌کپی و خلاقیت رعایت گردد.

قابلیت کنترل و شخصی‌سازی: در آینده سیستم‌ها باید تعامل‌پذیرتر شوند. ممکن است واسطهای کاربری نوین برای این منظور ابداع شود – مثلاً یک زبان میانی شبه‌موسیقایی که کاربر بتواند با آن به مدل بگوید "ملودی را غمگین‌تر کن" یا "این بخش را تکرار کن اما با پیانو". همچنین امکان یادگیری سلیقه‌ی شخصی کاربر می‌تواند مطرح شود؛ بدین صورت که مدل از طریق بازخورد کاربر در طول زمان بفهمد وی چه نوع خروجی‌هایی را بیشتر می‌پسند و خروجی‌های آتی را مطابق آن تنظیم کند.

ترکیب با ورودی‌های چندحسی: یک جهت جالب دیگر، چندوجهی‌تر شدن ورودی‌ها است. به عنوان نمونه، مدل‌هایی که همزمان تصویر و متن را می‌گیرند تا موسیقی بسازند (مثلاً تولید موسیقی متن یک فیلم بر اساس توضیحات صحنه و فریم‌های ویدئویی). یا حتی استفاده از سیگنال‌های زیستی (مانند ضربان قلب یا EEG) به همراه متن برای تولید موسیقی‌های درمانی کاملاً شخصی‌سازی‌شده. چارچوب‌های مولد چندمدلی (مثل MusicLM که ایده ترکیب ملودی صوتی و متن را داشت) احتمالاً گسترش خواهند یافت.

همکاری انسان و AI در آهنگسازی: دورنمایی که بسیاری به آن اشاره می‌کنند، استفاده از این مدل‌ها به عنوان ابزار کمکی آهنگسازان است نه الزاماً جایگزین آن‌ها. به عنوان مثال، آهنگساز ممکن است طرح کلی یک قطعه یا ملودی اصلی را خود بسازد، سپس از مدل بخواهد تنظیم (arrangement) آن را در سبک‌های مختلف امتحان کند یا بخش‌های هارمونی را پُر کند. بدین ترتیب، خلاقیت انسان و سرعت و مهارت ماشین تلفیق می‌شود. این امر مستلزم توسعه‌ی واسطها و قابلیت‌های خاص در مدل‌هاست (مثلاً ورودی چندلایه: ملودی انسان همراه با متن توضیح برای تنظیم). برخی مدل‌های کنونی تا حدی این قابلیت را نشان داده‌اند (MusicGen با ورودی ملودی راهنما).

جنبه‌های اخلاقی و حقوقی: در پایان باید یادآور شد که مانند سایر عرصه‌های تولید محتوای مصنوعی، این حوزه نیز با پرسش‌های اخلاقی روبروست. از جمله اینکه آیا استفاده از قطعات موسیقی موجود برای آموزش، نقض حقوق آن‌هاست یا خیر (بحث منصفانه بودن استفاده داده)، یا در صورت تولید یک قطعه خیلی شبیه به یک اثر معروف، تکلیف مالکیت معنوی چیست؟ همچنین امکان سوءاستفاده (مثلاً تولید موسیقی‌های حاوی پیام‌های تنفرآمیز) باید در نظر گرفته شود. خوشبختانه جامعه پژوهشی AI توجه جدی به این موارد دارد و انتظار می‌رود همراه با پیشرفت فنی، چارچوب‌های قانونی و اخلاقی مناسب نیز تدوین گردد.

به طور جمع‌بندی، تولید موسیقی مبتنی بر تحلیل احساسات و پردازش زبان طبیعی اکنون در مرحله‌ای قرار دارد که می‌توان آن را انقلابی در تعامل انسان–کامپیوتر در حوزه هنر دانست. کاربران قادرند ایده‌های ذهنی خود را به زبان بیان کنند و سیستم هوشمند آن را به زبان موسیقی ترجمه کند. هرچند هنوز راه زیادی تا رسیدن به ظرافت و عمق هنرمندان بزرگ باقیست، اما روند توسعه نشان می‌دهد که کیفیت و خلاقیت این مدل‌ها به سرعت در حال بهبود است. پیش‌بینی می‌شود در آینده‌ی نزدیک، این فناوری به صورت عمومی‌تر در دسترس آهنگسازان، بازی‌سازان، تولیدکنندگان محتوا و حتی افراد عادی قرار گیرد و فصل جدیدی در تلفیق هنر و هوش مصنوعی رقم بخورد.

مراجع

[1] Agostinelli, A., et al. (2023). “MusicLM: Generating Music From Text.” arXiv preprint arXiv:2301.11325.

[2] Davis, H., & Mohammad, S. (2014). “Generating Music from Literature.” Proc. of CLfL Workshop, EACL 2014.

[3] Dynamic game soundtrack generation in response to a continously varying emotional trajectory, Williams, D., Kirke, A., Eaton, J., Miranda, E. 11-13 Feb 2015, London, England.

[4] Ferreira, L. N., & Whitehead, J. (2019). “Learning to Generate Music with Sentiment.” ISMIR 2019.

[5] Kreuk, F., et al. (2023). “AudioGen: Textually Guided Audio Generation.” ICLR 2023

[6] Zhao, Y., et al. (2025). “AI-Enabled Text-to-Music Generation: A Comprehensive Review of Methods, Frameworks, and Future Directions.” Electronics, 14(6), 1197.

[7] Forsgren, S., & Martiros, H. (2022). “Riffusion: Stable Diffusion for Real-Time Music Generation.”

[8] MuLan: A Joint Embedding of Music Audio and Natural Language, Qingqing Huang, Aren Jansen, Joonseok Lee, Ravi Ganti, Judith Yue Li, Daniel P. W. Ellis, 26 Aug 2022

[9] copet, J., et al. (2023). “Simple and Controllable Music Generation.” (MusicGen model card).

[10] Huang, Z., et al. (2023). “Noise2Music: Text-conditioned music generation with diffusion models.”

[11] Schneider, A., et al. (2023). “Moûsai: Text-to-Music Generation with Long-Context Latent Diffusion.”

[12] Li, X., et al. (2024). “JEN-1: Text-Conditioned Universal Music Generation with Omnidirectional Diffusion Models.”

[13] Yang, Y., et al. (2022). “DiffSound: Discrete Diffusion Model for Text-to-sound Generation.”

[14] Liu, J., et al. (2023). “AudioLDM: Text-to-Audio Generation with Latent Diffusion Models.”

[15] Yoonjin Chung, Pilsun Eu, et 21 Feb 2025: “KAD: No More FAD! An Effective and Efficient Evaluation Metric for Audio Generation.”

[16] Mubert Inc. (2022). “Mubert: Text to Music Generator.” (Online API documentation).

[17] Agostinelli, A., et al. (2023). “MusicLM examples.” (Online samples) Google Research, https://google-research.github.io/seanet/musiclm/examples/
[18] https://en.wikipedia.org/wiki/Mean_opinion_score
[19] Gonzales, R., et al. (2024). “A Retrieval Augmented Approach for Text-to-Music Generation.”
[20] Huang, N., et al. (2024). “Aligning Text-to-Music Evaluation with Human Preferences.”
[21] https://huggingface.co/facebook/musicgen-small
[22] https://medium.com/@AIBites/musicgen-from-meta-ai-model-architecture-vector-quantization-and-model-conditining-explained-f9a030382f7d
[23] https://openlaboratory.ai/models/musicgen