تبدیل متن به گفتار با هوش مصنوعی

احتمالاً تا به امروز ویدئو های زیادی از دیپ فیک دیده‌اید که در آنها، افراد مشهوری نشان داده می‌شوند که درمورد موضوعی، دقیقاً با صدای خودشان و حرکات دقیق صورتشان صحبت می‌کنند. یکی از معروف‌ترین این ویدئو ها، ویدئو مورگان فریمن است که می‌توانید در این لینک (bit.ly/3ol2Ft9)مشاهده کنید. ویدئو های دیپ فیک از تکنولوژی‌های مختلف هوش مصنوعی استفاده می‌کنند؛ در این مقاله، تکنولوژی Text-to-Speech بررسی می‌شود که یکی از کاربردهای آن، استفاده در همین ویدئو های دیپ فیک می‌باشد.

حسین شعله رسا
دانشجو مهندسی کامپیوتر
دانشکدگان فارابی دانشگاه تهران
h.sholehrasa@ut.ac.ir

تکنولوژی Text-to-Speech (TTS)به کامپیوتر امکان تبدیل متن به گفتار را می‌دهد. در این فرایند، یک متن به‌عنوان ورودی در سیستم TTS قرار می‌گیرد؛ سپس سیستم با استفاده از الگوریتم‌های هوش مصنوعی، متن ورودی را به صوت تبدیل می‌کند.

یکی از جالب‌ترین نکات این تکنولوژی این است که سیستم قادر است تأکید، سرعت و بیان صحیح کلمات را به‌دقت اجرا کند. همچنین باتوجه‌به پیشرفت‌های جدید در حوزه TTS، صدای تولید شده توسط سیستم، بسیار نزدیک به صدای انسان و برای کاربران به‌راحتی قابل‌فهم است.

برای یادگیری نحوه تبدیل متن به صوت در هوش مصنوعی، مراحل زیر به طور مفصل بررسی می‌شوند:

۱. تجزیه‌وتحلیل متن: در این مرحله، متن ورودی به کمک پردازش زبان طبیعی (NLP) و تکنیک‌هایی مانند شبکه‌های عصبی بازگشتی (RNN) تحلیل شده و به‌صورت خودکار بخش‌های مختلف متن شامل کلمات و علامات نگارشی شناسایی و تجزیه می‌شوند.

به طور مثال در جمله «امروز هوا خیلی خوب است» اجزای مختلف جمله شامل فعل و فاعل جمله درک می‌شوند. همچنین، کلماتی که معمولاً همراه هم می‌آیند، مشخص می‌شوند. مانند کلمه «امروز» که نشان‌دهنده زمان است، همراه با «هوا» استفاده می‌شود. بعد از آن وضعیت هوا مانند «خوب» یا «بارانی» مشخص می‌شوند. هوش مصنوعی با نتیجه‌گیری این موارد با دانش خود می‌تواند بفهمد که این متن درمورد هوا است و مفهوم مثبتی دارد.

۲. تبدیل متن به صدا: بخش‌های تجزیه شده متن به بخش‌های کوچک‌تر تقسیم می‌شوند و به‌صورت خطی برای تولید صدا استفاده می‌گردند. این امر با استفاده از تکنولوژی‌های مختلفی از جمله ترکیب صداهای ماشینی و گفتار بشری، ترکیب فرکانس‌های مختلف صوت و تنظیم سرعت صدا انجام می‌شود.

در مرحله تولید صوت، ابتدا از پارامتری به نام مجموعه فونمی استفاده می شود. این مجموعه فونمی شامل تمامی واحدهای صوتی است که به‌عنوان اجزای سازندهٔ گفتار استفاده می‌شوند. بعد از آن، سیستم با استفاده از الگوریتم‌های پردازش سیگنال‌های صوتی و ترکیب فونم‌های مختلف، صدای متن را تولید می‌کند.

ازآنجاکه پردازش گفتار به‌دقت بالا و انعطاف‌پذیری کافی نیاز دارد، اغلب از شبکه‌های عصبی با ساختارهای مختلف مانند شبکه عصبی مولد برای تولید صدا استفاده می‌شود. این شبکه‌ها از یک مجموعه‌دادهٔ آموزشی برای یادگیری الگوهای صوتی بهره می‌برند و با استفاده از این الگوها ورودی متنی را به صدای گفتار تبدیل می‌کنند.

برای مثال، فرض کنید که متن ورودی «سلام، من یک نرم‌افزار برای تولید خودکار متن به صوت هستم» باشد؛ در اولین مرحله، این متن به بخش‌های کوچک‌تری مانند «سلام»، «من»، «یک نرم‌افزار»، «برای تولید خودکار متن به صوت» و «هستم» تقسیم می‌شود. این بخش‌ها سپس به ترتیب، تولید صدا می‌شوند تا یک خروجی صوتی به دست بیاید.

در مرحله بعدی، با تقسیم جمله به اجزای دستور زبانی آن (مرحله تحلیل)، تن صدا برای هر بخش مشخص می‌شود؛ کلمه «هستم.» به معنی پایان جمله می‌باشد و باید تن صدا از بالابه‌پایین باشد، و کلمه «سلام»، به این دلیل که بعد از آن ویرگول آمده است، باید با یک مکث کوتاه بیان شود.

۳. بهینه‌سازی کیفیت صدا: در این مرحله، برای بهبود کیفیت صدا شبکه های عصبی به کار می‌روند تا خطاهای موجود در صدای تولید شده را شناسایی و رفع کنند. ممکن است در خروجی صوتی، برخی از کلمات یا عبارات به‌درستی تلفظ نشوند یا کیفیت صدای تولید شده پایین باشد. با تحلیل دقیق داده‌های ورودی، بهبودهای لازم ایجاد شده و کیفیت صوت خروجی بهتر می‌شود.

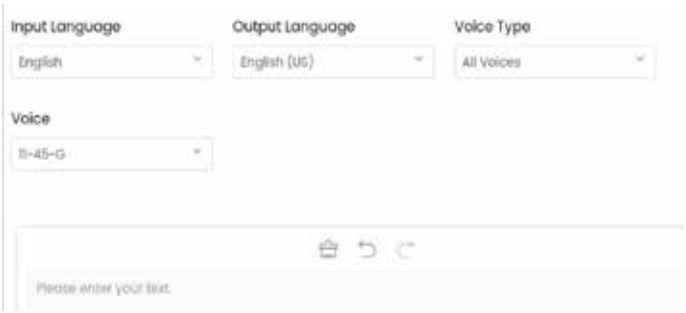
۴. تحلیل صوت: سیگنال صوتی، یک سیگنال الکتریکی است که توسط میکروفون یا سایر دستگاه‌های تبدیل صدا به سیگنال الکتریکی، ضبط می شود. در این سیگنال میزان قدرت صدا، فرکانس، مدت‌زمان و دیگر ویژگی های آن ذخیره می‌شوند.

تحلیل صوت به معنای بررسی و تحلیل ویژگی‌های صوتی است که در یک سیگنال صوتی وجود دارد. در ابتدا ویژگی‌های مختلفی مانند طول‌موج، فرکانس، قدرت و... از این امواج استخراج می‌شوند و سپس با استفاده از این ویژگی‌ها، الگوریتم‌های یادگیری ماشین می‌توانند الگوهای مختلفی را در سیگنال‌های صوتی تشخیص دهند. هدف نهایی این مرحله، تشخیص الگوهای صوتی یک فرد در بیان جملات مختلف است؛ تلفظ‌ها، لحن بیان و سرعت بیان کلمات در این فرد مشخص می‌شوند.

حال با انجام این فرایندها، هوش مصنوعی می‌تواند الگوهای بیانی یک فرد را یاد بگیرد و با دادن متن ورودی جدید، صوت جدیدی را تولید کند که بسیار شبیه به بیان آن فرد می‌باشد. باتوجه‌به متمایز بودن بیان هر فرد برای ساخت چنین هوش مصنوعی‌ای باید دقایقی از صدای صحبت‌کردن فرد موردنظر را استخراج کرد و مدل را بهبود داد تا این نوع از بیان را نیز هوش مصنوعی یاد بگیرد و به تفکیک از آن استفاده کند.

برای امتحان‌کردن این تکنولوژی، با ثبت‌نام در سایت‌های زیر می‌توانید با دادن متن دلخواه و انتخاب فرد موردنظر، صدای تولید شده آنها را بشنوید و قدرت حیرت‌انگیز این تکنولوژی را مشاهده کنید.

topmediai.com/text-to-speech
beta.elevenlabs.io



یکی از نکاتی که در این سایت‌ها می‌توانید مشاهده کنید این است که علاوه بر تولید سیگنال‌های صوتی مشابه فرد تعیین شده، می‌توان ویژگی های مختلف سیگنال صوتی مانند سرعت آن را تغییر داد و صداهای مشابه یا لحن‌های خاص‌تری با صدای موردنظر تهیه کرد.

با پیشرفت‌های بیشتر در حوزه‌های شناسایی صوتی و پردازش زبان، سیستم‌های TTS با هوش مصنوعی برای بسیاری از کاربردهای مفید در حوزه‌های گوناگون مانند آموزش، تبلیغات، پادکست، بازی‌های ویدئویی و بسیاری زمینه‌های دیگر قابل‌استفاده هستند. در نهایت، این حوزه به شرکت‌ها و کاربران عادی کمک می‌کند تا با یک تکنولوژی پیشرفته، متن ها را به گفتاری واضح و قابل‌فهم تبدیل کنند.