امکان سنجی ساخت

مدل جدید TTS در زبان فارسی

(کیفیت صدای طبیعی تر)

فاطمه وحیدیونسی

مرکز نوآوری علوم و فناوری های شناختی

دانشگاه علم و صنعت ایران

تاریخ تدوین: ۱۴۰۳/۸/۱۶

هدف این گزارش:

هدف از تدوین این گزارش، امکان سنجی و بررسی نیازمندی ها و منابع مورد نیاز جهت ساخت الگوریتم (و یا) مدل های جدید TTS در زبان فارسی است. مواردی که در زیر بررسی می شود، بررسی منابع، بودجه و زمان مورد نیاز برای اجرای این پروژه است. بدیهی است که با توجه به تحقیقاتی بودن زمینه پروژه، لزوما با هزینه کرد زمان و منابع ذکر شده، به نتایج بهتر دست پیدا نخواهد شد. تصمیم گیری در این مورد برعهده تیم همکار است.

۱- دیتاست مورد نیاز:

برای ساخت مدل های فعلی(با در نظر گرفتن بیان و کیفیت صدا در آنها) از حداقل ۱۰ ساعت صوت فاقد نویز که به طور حرفه ای و استودیویی ضبط گردیده استفاده شده که مشخصا برای هدف مدنظر این گزارش که کیفیت بالاتر است، به "حداقل" ۱۵ ساعت صوت با کیفیت نیاز است.

برای رسیدن به این مهم از دو راه می توان استفاده کرد:

الف) ساخت دیتاست از طریق ضبط صدا: این دیتاست می بایست شامل متن های transcript تک جمله ای به همراه صوت خوانش مربوط به هر یک از آنها باشد. همانطور که اشاره شد، صوت مدنظر باید به طور استودیویی، در فضای فاقد نویز، با دستگاه های ضبط حرفه ای مانند Zoom Recorder ضبط شده باشد. لحن بیان در همه صوت های ضبط شده می بایستی یکسان باشد. رعایت نکاتی مانند میزان فاصله تا میکروفون برای حفظ ویژگی های صوتی بسیار مهم است.

پس از ضبط، نیاز به بازبینی صوت ها و متن های مربوطه، توسط یک ناظر است، تا از حذف یا اضافه شدن ناخواسته کلمات در خوانش جلوگیری شود.

فایل های صوتی باید دارای این خصوصیات باشند:

- بدون خطا: هر گونه تپق یا خطا در گفتار فایلها باعث کاهش بازده میشود. کلمات باید دقیقا به صورتی که نوشته شده خوانده شوند. مثلا اگر کلمه "خواندند" در عبارت آمده بود نباید به صورت "خوندن" تلفظ بشود اما اگر در متن کلمه به صورت عامیانه وجود داشت، مثلا اگه به جای اگر باید به همان صورت عامیانه، "اگه" خوانده شود.
- نداشتن نویز زمینه: بازده کار به نویز زمینه به شدت حساس است، بنابراین می بایست از زیرساخت های ضبط حرفه ای استفاده شود.
- لحن ثابت: به هیچ عنوان نباید لحن متفاوت برای جملات متفاوت استفاده کرد. هر خط از متن باید به صورت مستقل و بدون در نظر گرفتن خطوط قبل و بعد اجرا شود. لحن صحبت فارغ از موضوع و این که نقل قول از چه کسی است باید با لحن ثابت و شخصیت ثابت اجرا شود به طوری که اگر یک جمله در دو قسمت از فایل تکرار شود صدای خروجی باید بیشترین شباهت ممکن به یکدیگر را داشته باشد. صوت های ضبط شده باید از نظر احساسی خنثی باشند. مثلا جملات غمگین نباید با لحن غمگین خوانده شوند. در طول ضبط گوینده باید یک شخصیت ثابت داشته باشد و تمام جملات باید متناسب با آن شخصیت خوانده شود. مثلا ممکن است شخصیتی سرزنده انتخاب شود. پس در تمام طول ضبط باید این سرزندگی و خوشبینی حفظ شود و ثابت بماند. به خاطر رسیدن به ثبات هر چه شخصیت صوتی به حالت عادی گوینده نزدیک تر باشد نتیجه بهتر خواهد بود.

- صدای طبیعی: علاوه بر نکته قبلی باید به این نکته هم توجه داشت که اجرای صوتی نباید خشک و با لحن ماشینی باشد. هر چقدر اجرا و صدا طبیعی تر باشد، خروجی بهتر خواهد بود. مطلوب ترین سرعت گفتار سرعت طبیعی است. نه به طور آزار دهنده ای کند و نه به صورت غیر قابل فهمی تند. در مورد تن صدا و گام هم به همین منوال در ضمن به علایم نگارشی در جمله هم حتما می بایست توجه شود. جملات سوالی یا ویرگول ها در جمله حتما با دقت لازم اجرا شوند. در کل مهمترین خصوصیت اجرا طبیعی بودن و واضح بودن در عین ثبات در طول تمام فایل ها است.

ب) استفاده از صوت های ضبط شده با کیفیت مناسب: در این روش می توان از صوت های ضبط شده با کیفیت توسط یک فرد و در دسترس در شبکه های اجتماعی استفاده کرد. این روش فرآیند ساخت دیتاست را ساده تر و سریعتر می کند اما نیاز به کسب مجوز از مالک صدا و همچنین پیاده سازی صوتی آن است. به عنوان مثال می توان از صوت پادکست ها استفاده نمود. پیاده سازی متن به ۳ حالت قابل اجرا است:

- پیاده سازی متن صوت ها توسط یک اپراتور
- پیاده سازی متن توسط مدل های STT(صوت به متن) و سپس اصلاح آنها(به علت خطاهای موجود در این مدل ها و همچنین اضافه کردن علایم نگارشی به متن که در خروجی این مدل ها وجود ندارد.)
- درخواست متن پادکست از مالک پادکست و سپس بازبینی متن توسط ناظر
- استفاده از مجله های صوتی یا پادکست هایی که به همراه صوت، متن پادکست را نیز منتشر کرده اند.

همان طور که اشاره شد، جز پادکست، استفاده از مجله های صوتی نیز در این حالت پیشنهاد می شود. البته مدل های فعلی مانند Piper نیز دقیقا بر روی همین صداها گسترش یافته اند، به همین خاطر تغییر دیتاست لزوما تاثیری در بهبود خروجی مدل نخواهد داشت. تنها کمکی که تغییر دیتاست به ما می کند، در تغییر صدای گوینده(و یا گویش یا لحنی دیگر) است، مثلا استفاده از صدایی سر زنده تر.

۲- فرآیند Preprocessing:

پس از ساخت دیتاست و انجام بازبینی توسط ناظر، نیاز به انجام فرآیند Preprocessing بر روی صوت ها و متن هاست. نیاز است متن ها در قالب یک فایل csv ذخیره شوند. صوت ها می بایست تک کاناله شده(mono) و درsample rate پایین تر ذخیره شوند. هر جمله می بایست در قالب یک فایل صوتی به طور جداگانه ذخیره شود(استفاده از مدل های کشف سکوت).

۳- فرآیند مدلسازی:

جهت رسیدن به کیفیتی بهتر از کیفیت مدل های فعلی، نیاز به بهبود مدل است. در این مرحله پیاده سازی مدل های به روز در مقالات که تاکنون در TTS فارسی پیاده نشده اند اما نمونه های آنها در زبان های دیگر کیفیت مطلوب را داراست، قدم اول است. در صورت عدم پاسخ دهی مناسب این مدل ها، نیاز به انجام تحقیقات جهت ساخت الگوریتم جدید جهت تست کیفیت بهتر است. باید در نظر داشت که لزوما یک الگوریتم جدید نمی تواند کیفیت بالاتری را به دست دهد. همانطور که برخی الگوریتم های جدید پیاده شده در زبان انگلیسی نتوانسته اند کیفیتی فراتر از کیفیت مدل هایی مانند piper دست پیدا کنند. با توجه به زمان و بودجه مورد نیاز برای این امر، لذا هزینه فرصت این فرآیند باید در تصمیم گیری ها لحاظ شود.

۴- زیرساخت های مورد نیاز:

مدل های TTS فعلی اغلب بر روی GeForce 4090 و سخت افزارهای مشابه آن توسعه یافته و آموزش دیده اند. پلن رایگان GPU که در پلتفرم Colab در دسترس است نیز برای مدلسازی پروژه های TTS قابل استفاده است اما سرعت آموزش و تست را کاهش می دهد.

۵- زمان مورد نیاز جهت اجرای این فرآیند:

زمان مورد نیاز به دو فاکتور، نحوه ساخت دیتاست و پاسخ دهی یا عدم پاسخ دهی مدل های موجود در مقالات بستگی دارد.

جهت ساخت دیتاست با روش اول پیش بینی می شود حدود سه ماه زمان برای یافتن گوینده و استودیو، انتخاب متن، ضبط صدا و بازنگری در آن و انجام عملیات پیش پردازش بر روی صدا نیاز است. درصورت استفاده از روش دوم، این زمان به یک ماه قابل تقلیل است اما ممکن است صدا با لحن مد نظر در روش دوم یافت نشود.

جهت انجام پژوهش و تحقیق و اجرای مدل های فعلی موجود در مقالات که تاکنون مورد پیاده سازی در زبان فارسی قرار نگرفته اند، حدود سه ماه زمان نیاز است. در صورت عدم بازدهی مدل های مذکور و نیاز به انجام تحقیقات بیشتر جهت ساخت الگوریتمی جدید، نیاز به جذب نیروی انسانی در زمینه صوت و حدود یکسال زمان جهت تحقیق، پیاده سازی و تست الگوریتم جدید است.

با توجه به اینکه ساخت دیتاست و انجام تحقیقات به طور موازی قابل پیگیری است، در حالت اول(اجرای مدل های موجودِ پیاده نشده در زبان فارسی) حدود سه ماه و در حالت دوم حدود یکسال زمان نیاز است. تاکید می شود که مدل های جدید توسعه داده شدهء احتمالی، لزوما کیفیت بهتری را تضمین نمی کنند.

۶- منابع مورد نیاز:

منابع مورد نیاز به دو فاکتور، نحوه ساخت دیتاست و پاسخ دهی یا عدم پاسخ دهی مدل های موجود در مقالات بستگی دارد.

- بودجه: بودجه مورد نیاز جهت اشتراک GPU از سایت هایی مانند Runpod، جذب گوینده و ضبط صدا(درصورت ساخت دیتاست)، جذب نیروی انسانی(در حالت دوم- ساخت الگوریتم جدید) و سیستم و سخت افزار جهت تحقیقات و اجرای مدل ها.
- سخت افزار: در بخش "زیرساخت های مورد نیاز" بررسی شد.
- منابع انسانی: در حالت دوم در توسعه مدل ها(ساخت الگوریتم جدید) نیاز به حداقل یک نیروی انسانی متخصص هوش مصنوعی در زمینه صوت است. در حالت ساخت دیتاست، نیاز به یک گوینده و یک ناظر جهت بررسی و تصحیح خطاهای خوانشی است.
- دیتاست: در بخش "دیتاست مورد نیاز" توضیح داده شد.