رعایت نکات زیر جهت ساخت دیتاست برای آموزش یک مدل TTS(Text To Speech) حایز اهمیت هستند:

۱- **مرحله جمع آوری صوت از صدای مدنظر:**

جمع آوری ۱۵ ساعت صوت خام از صدای مدنظر. منظور از صوت خام، صوتی است که دارای شرایط زیر باشد:

- نداشتن نویز زمینه: بازده و کیفیت مدل به نویز زمینه به شدت حساس است، بنابراین می بایست هرگونه نویز از صدا حذف شود.
- حذف صدای زمینه: صدای جا به جایی اشیا، صدای دستگاه های الکتریکی در محل ضبط، صدای سرفه، صاف کردن صدا توسط گوینده، حتی نفس گرفتن وی بین صحبت ها می بایستی حذف شود. درواقع هیچ گونه صدایی جز بیان کلمات توسط گوینده نباید در صوت خروجی وجود داشته باشد.
- حذف اکو: معمولا ناشی از ضبط در سالن یا استفاده از میکروفونِ تنظیم نشده است که می بایست از زمینه صوت حذف شود.
- لحن ثابت: گوینده در بخش های مختلف گفت و گو ممکن است لحن های متفاوت(خوشحال، عصبانی یا ..) داشته باشد. در صوت نهایی تنها باید لحن اصلی(طبیعی) گوینده وجود داشته باشد. باید بخشی از صوت که در آن گوینده می خندد یا با خنده صحبت خود را بیان می کند و یا با عصبانیت/خشم/ ناراحتی(هر حالتی خارج از حالت طبیعی گفت و گو) صحبت می کند از صوت نهایی حذف شود. به هیچ عنوان نباید لحن متفاوت برای جملات متفاوت وجود داشته باشد، به طوری که اگر یک جمله در دو قسمت از فایل تکرار شود صدا و لحن آنها باید بیشترین شباهت ممکن به یکدیگر را داشته باشد. در کل مهمترین خصوصیت صوت ها، طبیعی بودن و واضح بودن در عین ثبات در طول تمام فایل ها است.
- ترجیحا فایل های صوتی می بایست در فرمت wav "دانلود" و "ذخیره" شوند.
- با توجه به اینکه زبان مدنظر ما در این پروژه برای پیاده سازی، زبان عربی است و قرار است NPC با زبان عربی سخن بگوید، تنها فایل های صوتی که در آن گوینده به زبان عربی صحبت می کند قابل استفاده خواهد بود.

۲- **ساخت متن Transcript از صوت:**

پس از ایجاد فایل صوتی با شرایط فوق، فایل با فرمت csv از transcript(متن متناسب با هر جمله در فایل صوتی) ایجاد می شود.

جهت ساخت این فایل می توان از زیرنویس تولیدی یوتیوب یا مدل های STT و VAD استفاده کرد و سپس خروجی این روش ها را از نظر موارد زیر مورد بازبینی قرار داد:

- بدون خطای نگارشی و املایی: هر گونه خطا در نوشتارِ پیاده سازی شده از صوت ها باعث کاهش بازده می شود. کلمات باید دقیقا به صورتی که بیان شده نوشته شده باشند. مثلا اگر کلمه " خوندن" در فایل صوتی آمده است نباید متن آن به صورت "خواندند" پیاده سازی شده باشد.
- علامت سوال، نقطه انتهای جمله یا ویرگول ها در جمله حتما با دقت باید پیاده سازی شده باشند.
- اعراب گذاری ها به درستی پیاده سازی شود.
- فایل csv باید شامل ۲ ستون(نام فایل صوتی مربوطه، متن پیاده سازی آن) باشد. (لزومی ندارد هر عبارت شامل یک جمله باشد. درصورت استفاده از مدل های STT و VAD جهت پیاده سازی متن، ممکن است هر عبارت، شامل نیمی از یک جمله یا حتی چند جمله باشد). (مثال در تصویر زیر)

ID	Text	Audio File
۱	ما از هم پیمانان خود حمایت می کنیم.	1.wav
۲	افتخار ما، شهادت است.	2.wav
۳	به وعده های خود عمل خواهیم کرد.	3.wav
۴	از برادران خود سپاسگزاریم.	4.wav

شکل ۱- نمونه یک فایل csv از transcriptها

۳- ادامه فرآیند Preprocessing صوت:

پس از ساخت دیتاست و انجام بازبینی توسط ناظر، نیاز به انجام فرآیند Preprocessing بر روی صوت هاست. صوت ها می بایست تک کاناله شده(mono) و در22,050 Hz sample rate ذخیره شوند.(صوت های قبل از این پردازش نیز جهت اطمینان ذخیره شوند). هر عبارت می بایست در قالب یک فایل صوتی به طور جداگانه، در فرمت wav ، ذخیره شود(می توان از مدل های کشف سکوت یا VAD استفاده کرد).