طرح پیشنهادی پروژه

دستیابی به دانش و ساخت سیستم تولید گفتار انسانی
با استفاده از متدهای سنتز صوت و مدل مبدل متن به صوت

پیشنهاد دهنده‌ی طرح

فاطمه وحیدیونسی

تاریخ ارائه طرح

۳۱ اردیبهشت ۱۴۰۴

بسمه تعالی

۱- کلیات طرح:

۱-۱ عنوان طرح پیشنهادی:

عنوان به زبان فارسی:

دستیابی به دانش و ساخت سیستم تولید گفتار انسانی با استفاده از متدهای سنتز صوت و مدل مبدل متن به صوت

عنوان به زبان انگلیسی:

Acquiring knowledge and building a human speech production system using voice synthesis methods and text-to-speech converter models

۱-۲ مدت اجرا (ماه): ۶ماه (برای سنتز و ساخت ۱۶ صوت موجود در جدول صفحات ۹-۱۰-۱۱)

۱-۳ اعتبار مورد نیاز : ۱۱۰۰ (میلیون ریال)

۲- مشخصات مجری/ همکاران اصلی طرح:

۲-۱) مشخصات مدیر و همکاران پروژه:

ردیف	نام و نام خانوادگی	آخرین مدرک تحصیلی	رشته/گرایش	نام موسسه محل تحصیل	سال دریافت مدرک	نقش در پروژه	فعالیت مرتبط بر اساس گانت پروژه	تلفن همراه	کد ملی
۱	فاطمه وحیدیونسی	کارشناسی ارشد	MBA	صنعتی شریف	۱۴۰۰	کارشناس ارشد(مجری)	مجری در تمامی مراحل گانت	۰۹۱۲۹۳۴۶۷۱۴	۰۰۱۹۲۱۷۷۳۰

۲-۲) خلاصه سوابق علمی-تحقیقاتی اعضای پروژه(طبق لیست بند ۲-۱ تکمیل شود):

ردیف	نام و نام خانوادگی	از تاریخ	تا تاریخ	موسسه ای که در این مدت با آن همکاری داشته اید	عناوین پروژه‌های پایان یافته و یا در حال اجرا را به همراه نوع(دانشی/فناوری/امکان سنجی/سامانه ای) و وضعیت(خاتمه یافته/جاری) آن بیان کنید.
۱	فاطمه وحیدیونسی	آبان ۱۴۰۲	اکنون	مرکز علوم شناختی دانشگاه علم و صنعت	توسعه و پیاده سازی مدل های هوش مصنوعی تولید متن از گفتار(ASR/STT)، تولید گفتار از متن(TTS) = در حال اجرا- فناوری+سامانه ای Voice Cloning، سنتز و پردازش صوت، تشخیص محتوای فایل صوتی/ویدئویی، استخراج کلیدواژه از متن فارسی، مدلسازی تشخیص احساسات از متون توییت عربی و صوت، تولید پاسخ در چت بات های Interactive با استفاده از مدل های LLM = خاتمه یافته-دانشی+ فناوری+سامانه ای
		اردیبهشت ۱۴۰۲	اردیبهشت ۱۴۰۳	شرکت پویانمای شمس	ساخت صدای شاهد و صدای دوبله با استفاده از مدلسازی هوش مصنوعی با تکنولوژی سنتز گفتار و TTS. ساخت مدل های یادگیرنده و مقلد صدای اشخاصِ نامی، جهت ساخت محتوا با استفاده از مدلسازی Voice Cloning= خاتمه یافته- فناوری

۳- اهداف پروژه:

توسعه سیستم مولد صوت(گفتار، صدای طبیعی انسان) با استفاده از پردازش ویژگی های صوتیِ یک سیگنال صوتی مبدا، همچنین ساخت دیتاست و آموزش مدل های TTS

اهداف دیگر:

- یافتن روش های علمی مختلف و متدهای پردازش سیگنال صوتی، جهت بازتولید صوت گفتار انسان از صوتی دیگر
- ساخت دیتاست و مدلسازی صوت اشخاص و شخصیت های خاص
- ساخت Interactive NPC هایی با قابلیت خواندن اشعار در کنار قابلیت خواندن نثر و متن
- ساخت Interactive NPC هایی چند زبانه(فارسی، انگلیسی و عربی)
- بازسازی صدای مشاهیر غیر معاصر نظیر حافظ، فردوسی و سعدی
- دستیابی به دانش توسعه و تولید سیگنال های صوتی انسان

٤- بیان مسأله و ضرورت اجرای پروژه (فوریت حل مسئله):

در دنیای فناوری صوتی، یکی از چالش های اصلی، امکان تولید تن یا لحن انسانی از طریق کنترل فرکانس های صوتی است. لازم به ذکر است که در این پروژه، هر جا از تولید صوت یا گفتار صحبت شده، منظور تولید تن یا لحن انسانی است. این پرسش مطرح است که آیا می توان با داشتن یک محدوده مشخص از فرکانس های صوتی، تن و لحن های مختلفی در گفتار را تولید کرد؟ این پروژه به دنبال توسعه یک سیستم پردازش صوتی است که بتواند با استفاده از تغییر فرکانس های صوتی، یا روش های پردازشی دیگر مانند استفاده از متد های TTS، لحن ها و تن های متنوع انسانی را تولید کند. این شامل مدل سازی فرایندهای گفتاری، تحلیل ویژگی های فرکانسی و ایجاد الگوهای صوتی با کیفیت بالا برای کاربردهای مختلف مانند شخصیت های مجازی، بازی های ویدیویی، دوبله و تولید محتوای صوتی سفارشی است. علاوه بر این، سیستم می تواند صداهای تولیدی را برای شخصیت های مختلف سفارشی سازی کرده و حتی صدای شخصیت های معروف را شبیه سازی کند تا در حوزه هایی نظیر دوبله، تولید محتوای صوتی و توسعه شخصیتهای مجازی کاربرد داشته باشد . در بسیاری از کاربردها، مانند موسیقی، بازی های ویدیویی، واقعیت مجازی و حتی تولید محتواهای آموزشی، نیاز به تولید صداهایی با تنوع بالا و قابلیت تنظیم دقیق احساس می شود. هدف از این پروژه، بررسی و پیاده سازی راهکاری برای تولید لحن و تن انسانی با قابلیت کنترل طیف های متنوع صوتی است که بتواند در حوزه های مختلف کاربردی مورد استفاده قرار گیرد .

۴-۱- چالش‌های پروژه:

- تولید صدایی طبیعی و مشابه گفتار انسانی
- تولید صوت هایی متعدد و متفاوت از "یک" صوت ورودی
- ساخت دیتاست صوت/متن با کیفیت بالا برای مدلسازی
- عدم وجود دیتای صوتی از برخی شخصیت های خاص مدنظر
- استفاده از روش های مختلف سنتز صوت در ایجاد صوت جدید
- امکان سنجی تغییر Timbre یا رنگ صوت در تولید صدای جدید
- بهینه سازی زمان اجرا و Inference

۵- شرح نیازهای کاربر (RFP):

این طرح براساس نیاز ساتخ/ مرکز نوآوری علوم و فناوری های شناختی دانشگاه علم و صنعت و به منظور به کار گیری در پروژه/سامانه "شخصیت‌های هوشمند غیربازیکن" تعریف و خدمات زیر مورد درخواست رده کاربر می باشد.

۶- ضرورت اجرای پروژه:

- **نیاز به صوت های متعدد برای کاراکترهای NPC متفاوت:** با توجه به تعدد کاراکترهای NPC، نیاز به تولید صدای گفتار انسانی متعدد و متنوع برای هر یک از این کاراکترهاست. به خصوص برای کاراکتر مربوط به شخصیت های خاص، نیاز به مشابه بودن صوت تولید شده با صوت این اشخاص وجود دارد، بنابراین نیاز است تا صدای این افراد جهت اجرای بهتر بازی، شبیه سازی شوند.
- **ساخت صدای شخصی سازی شده:** با استفاده از روش های مختلف مدلسازی TTS که در بخش های بعدی مفصلا توضیح داده شده است، امکان تولید صدای اشخاص و افراد خاص بوجود می آید. به علاوه در برخی نمونه ها، با استفاده از روش سنتز صوت، می توان بر اساس تغییر ویژگی های صوتی، صدایی با سن و جنسیت خاص و مدنظر تولید کرد.
- **کاربردهای تجاری و آموزشی:** همانطور که گفته شد توسعه سیستم مولد صوت انسانی، می تواند در رشته های مختلف از جمله ساخت NPCهای هوشمند، بازی های ویدیویی، شخصیت های مجازی، تولید صوت دوبله، تولید محتوای صوتی در تبلیغات و بازاریابی، تولید محتواهای آموزشی، تمرین‌های تعاملی و سناریوهای پیچیده در حوزه‌هایی مانند آموزش زبان و مدیریت بحران به کار آید.

- **همگام‌سازی با پیشرفت‌های جهانی:** جلوگیری از عقب‌ماندگی در رقابت جهانی از طریق نوآوری و بهره‌گیری از فناوری‌های نوین هوش مصنوعی.
- **تقویت فناوری‌های بومی:** ایجاد فرصت‌های راهبردی برای توسعه فناوری‌های بومی و گسترش کاربردهای آن در سطح جهانی.
- **مزایای روش "سنتز صوت":**

در این پروژه از سه روش جهت ساخت صوت گفتار انسانی استفاده می شود. سنتز صوت، مدلسازی TTS، مدلسازی ترکیبی TTS و Voice Cloning. در زیر مزایای ناشی از روش "سنتز صوت" بیان شده است:

- **عدم نیاز و وابستگی به دیتاست در ساخت صوت جدید:** در روش سنتز صوت، خروجی بر

اساس متدهای پردازش سیگنال های صوتی تولید شده و بنابراین به دادهء ورودی برای مدلسازی صوت نیاز ندارد. درحالیکه مدل های TTS وابسته به وجود دیتای صوتی با زمان بندی خاص برای هر صوت هستند.

- **افزایش سرعت و کاهش زمان مورد نیاز جهت ساخت صدای انسانی در کاربردهای مختلف:** با توجه به اینکه راه حل دیگر در تولید صدای انسانی، استفاده از مدل های TTS است که نیاز به ساخت دیتاستی از صوت و متن به شکل Align شده و دقیق دارد که امری زمان بر است، روش "سنتز صوت" می تواند زمان مورد نیاز برای ساخت اصوات جدید را تا ۱۰٪ زمان مورد نیاز برای پروژه های TTS کاهش دهد(۹۰٪ بهینه سازی زمان). این کاهش زمان مربوط به مرحله توسعه است، بدیهی است که در مرحله اجرا(در زمان اجرای بازی)، این مرحله حدود ۱۰٪ زمان بیشتری نسبت به روش TTS می طلبد، چرا که برای اجرای این روش، نیاز به اجرای مدل TTS و سنتز صوت با یکدیگر است که زمان بیشتری نسبت به اجرای یک مدل TTS به تنهایی نیاز دارد.
- **کاهش هزینه های سخت افزاری و نیروی انسانی:** همانطور که گفته شد، سیستم مولد صوت تا ۹۰٪ زمان تولید صوت های جدید انسانی را save(ذخیره سازی) می کند، در نتیجه با توجه به کاهش زمان اجرا و از طرف دیگر عدم نیاز به ساخت دیتاست برای تولید هر صوت، تعداد نیروی انسانی مورد نیاز برای اجرای پروژه را کاهش داده و به این دو طریق باعث کاهش هزینه های تولید صدای جدید می شود.

۷- سوالات و جنبه های نوآوری پروژه:

سؤالات پروژه

۱. درمورد اشخاصی که از آنها صدایی در زبان مدنظر وجود ندارد و یا حجم صدای موجود بسیار اندک است و برای مدلسازی TTS کافی نیست، می توان با روش Voice Cloning اقدام به ساخت دیتاست جدید و سپس مدلسازی نمود؟ کیفیت خروجی در این حالت به چه شکل خواهد بود؟
۲. آیا ساخت سیستم سنتز گفتار جهت ساخت اصوات جدید از صوت های موجود، آن طور که فرض شده، سبب کاهش زمان "توسعه" پروژه می شود؟
۳. استفاده از روش سنتز گفتار در "مرحله استفاده و Inference" به چه میزان سبب افزایش زمان اجرا می شود؟
۴. از طریق چه سنتز و پردازش هایی امکان ساخت صدایی طبیعی انسان وجود خواهد داشت؟
۵. هر کدام از این پردازش ها در چه محدوده مقادیری، صدای طبیعی تولید خواهند کرد؟
۶. با استفاده از این سیستم، از هر صوت ورودی، امکان ساخت چند صوت متفاوت و جدید وجود دارد؟

جنبه‌های نوآوری پروژه

- استفاده از روشی نوآورانه و جدید (استفاده از متد Voice Cloning) برای ساخت دیتاست صوتی، درمورد اشخاصی که از آنها صدایی در زبان مدنظر وجود ندارد و یا حجم صدای موجود بسیار اندک است.
- ساخت صدای اشخاص در زبانی به جز زبان مورد استفاده آنها(به عنوان مثال ساخت گفتار به زبان انگلیسی برای شهید سید حسن نصرالله)
- شبیه سازی و ساخت صدای اشخاصی که صوتی از آنها در دسترس نیست و یا صوت موجود بسیار اندک و کم کیفیت است.
- ساخت دیتاست هایی جدید برای برخی صوت های مدنظر
- استفاده از ادغام و ترکیب متدهای پردازش سیگنال های صوتی جهت ساخت صوت جدید
- ساخت سیستم با امکان شخصی سازی صوت تولیدی بر اساس ویژگی های "از پیش مشخص شده"
- ساخت صوت انسانی جدید، بدون مدلسازی و استفاده از تکنولوژی TTS و بدون نیاز به ساخت دیتاست(روش سنتز صوت)
- اجرای Real-time یا در لحظه صوت های سنتز شده جهت بهبود سرعت اجرای بازی

۸- مشخصات عمومی و فنی پروژه (اطلاعات کمی و کیفی پروژه)

ساخت صوت گفتار انسانی یکی از پیچیده ترین وظایفی است که هوش مصنوعی از عهده آن بر آمده است. هر یک ثانیه صوت تولید شده شامل ۴۴۱۰۰ فریم صوت است درحالیکه در مدل های تولید تصویر تنها یک فریم(تصویر) تولید می شود. حتی در مدل های تولید ویدئو در هر ثانیه حدود ۲۴ فریم تولید می شود که سبب تفاوت چشمگیر در میزان پیچیدگی مدل های تولید این خروجی ها دارد. جهت نیل به این هدف، ۳ روش کلی وجود دارد:

۱- در حالتیکه نیاز به تولید صدای شخصی خاص وجود دارد و یا نیاز به مشخصات خاصی در صدا مانند لحن، تنین، بیان و سن یا جنسیتی خاص است، می توان از مدل های TTS(Text To Speech) استفاده کرد. درصورتیکه از صدای مدنظر دیتای صوتی کافی(حداقل ۱۰ ساعت) موجود باشد، از این روش استفاده می کنیم. این روش دارای ۵ مرحله اصلی است:

- ساخت و پردازش دیتای صوتی(Audio Data) از صدای مدنظر

- ساخت متن Transcript از دیتای صوتی
- ویراستاری متن Transcript بر اساس Audio Data
- مدلسازی TTS
- ارزیابی کیفیت صوت خروجی و در صورت نیاز تغییر دیتای صوتی و اجرای مراحل بعدی

۲- مدلسازی TTS درمورد صدایی خاص اما بدون وجود Audio Data:

درصورتیکه نیاز به تولید صدای فردی خاص یا صدایی با مشخصه ای خاص(بیان و لحن خاص) را داشته باشیم اما دیتای صوتی از این صدا در زبان مدنظر موجود نباشد و یا دیتای موجود کافی نباشد(زیر ۱۰ ساعت) می توانیم از روش Voice Cloning برای ساخت دیتاست و سپس مدلسازی TTS استفاده کنیم. این روش برای اولین بار در حال استفاده است و کیفیت خروجی پس از تست مشخص می شود اما پیش بینی می شود که به کیفیتی قابل قبول (متناسب با حداقل صوت موجود) برسیم. در این روش پس از مشخص کردن زبان مدنظر، یک دیتاست صوتی،(از صدای هر شخصی، ترجیحا صدایی با لحن و بیان مشابه) ساخته شده، سپس با استفاده از مدل های Voice Cloning، صوت موجود در این دیتاست به صوت شخص مدنظر مبدل می شود. متن بدون تغییر باقی می ماند و در نتیجه در نهایت یک دیتاست جدید با صدای شخص مدنظر را خواهیم داشت که برای به عنوان ورودی برای مدلسازی TTS استفاده می شود.

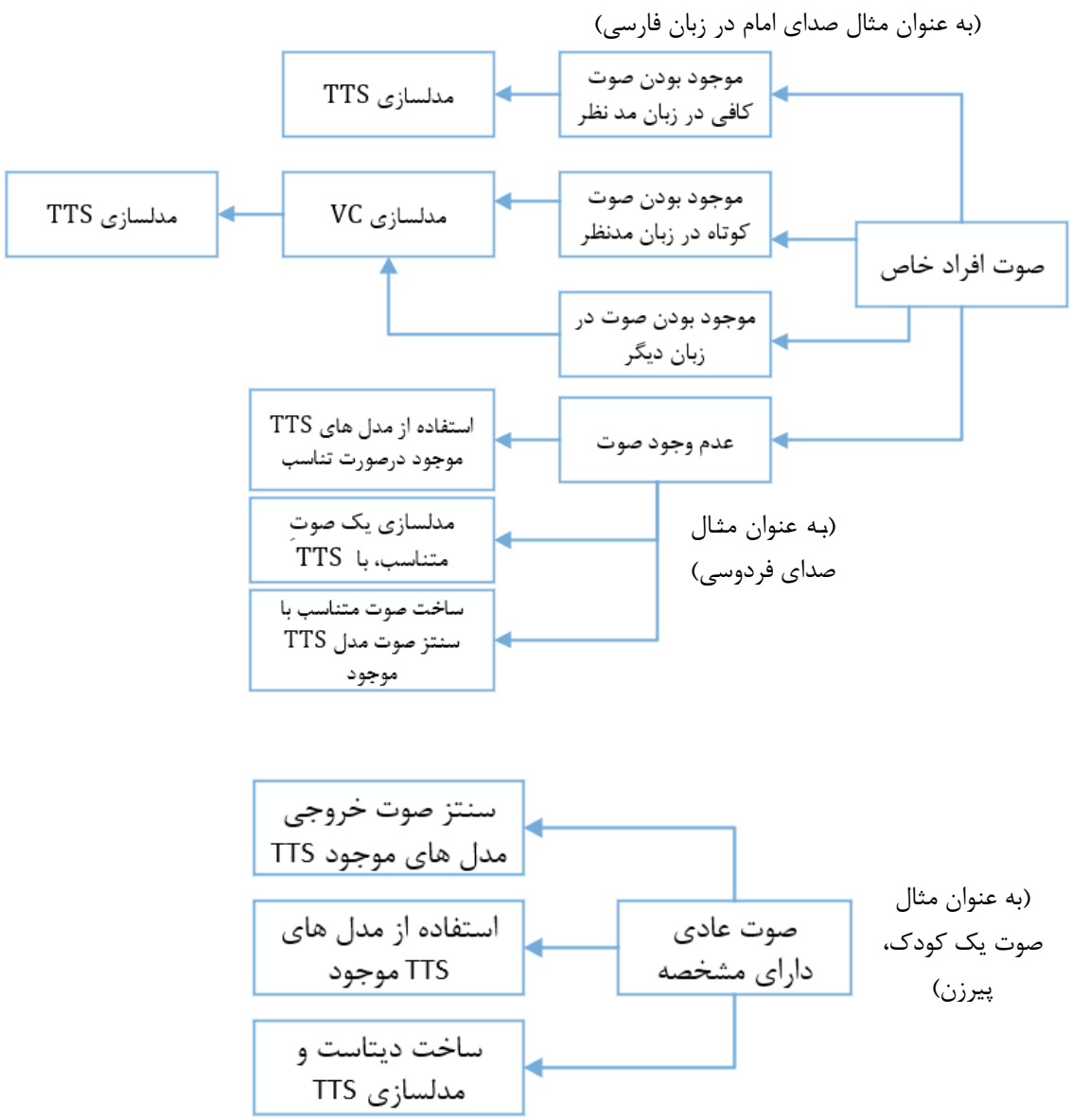
۳- تولید صوت با استفاده از پردازش سیگنال های صوتی:

راه حل سوم برای ساخت صوت انسانی، استفاده از روش های سنتز و پردازش ویژگی های سیگنال صوتی است. این روش نمی تواند یک صدای خاص(از شخصیتی خاص) را تولید کند، اما زمانیکه هدف، تولید صداهایی متنوع و متعدد باشد، کاربردی است. این روش به دلیل عدم نیاز به ساخت دیتاست و طی مراحل ذکر شده در بالا، سریعتر و کم هزینه تر است.

در این روش پیش بینی می شود، با استفاده از تغییر ویژگی های Pitch[1] و Speed در صوت های در دسترس، بتوانیم از هر صوت مبدا، دو صوت جدید را تولید کنیم. در کنار این ویژگی، به بررسی امکان تغییر Timbre[2] یا رنگ صوت(که یکی از ویژگی های اساسی در تفاوت صدای انسان هاست) می پردازیم. برای توسعه این کار از ساخت ماژول با زبان پایتون بهره می بریم.

تغییر ویژگی های دیگر صوتی مانند هارمونی، ادغام فرکانس ها و Formant ها می تولند به کلی کلمه مدنظر را تغییر دهد و بنابراین منتج به هدف این پروژه نمی شود. بنابراین با توجه به هدف این پروژه که ساخت لحن های جدید از یک صوت مبدا است، از سنتز pitch و سرعت بیان صوت، استفاده خواهیم کرد.

در زیر خلاصه ی سه روش فوق به همراه موارد استفاده آنها ترسیم شده است:



جدول زیر اهداف مربوط به این پروژه را همراه با متد مورد اســتفاده و نیازمندی های هر روش(مانند دیتاســت مورد نیاز) را نشان می دهد.

اشخاص	دسته بندی صدا	متد مورد استفاده	دیتاست مورد نیاز	نحوه جمع آوری دیتاست
امام خمینی	موجود بودن صوت کافی در زبان فارسی	مدلسازی TTS	ساخت دیتاست متن/صوت از سخنرانی های امام	نیاز به ۱۰ ساعت صوت "با کیفیت" است. می توان از صوت مصاحبه ها/سخنرانی های موجود استفاده کرد.

رهبر	موجود بودن صوت کافی در فارسی و عربی	مدلسازی TTS	ساخت دیتاست متن/صوت از سخنرانی های رهبر	نیاز به ۱۰ ساعت صوت "با کیفیت" است. می توان از صوت سخنرانی های موجود در سایت ایشان استفاده کرد.
شهید محمدباقرصدر	عدم وجود صوت با کیفیت	استفاده از مدل "احمد گنجی"	بدون نیاز به دیتاست	بدون نیاز به دیتاست
امام موسی صدر	عدم وجود صوت با کیفیت	سنتز خروجی مدل "احمد گنجی"	بدون نیاز به دیتاست	بدون نیاز به دیتاست
شهید مطهری	عدم وجود صوت کافی	مدلسازی ترکیبی VC+TTS ❋	ساخت یک دیتاست حداقل ۲۰ دقیقه ای از صدای ایشان	حداقل ۲۰ دقیقه "با کیفیت" صوت "مصاحبه ها" در اولویت است.
فردوسی	عدم وجود صوت	مدلسازی TTS	۱:پنج ساعت دیتاست متن ۲: پنج ساعت دیتاست از شعر به همراه صوت خوانش.	می توان از سایت گنجور برای ساخت دیتاست استفاده کرد[1]
حافظ	عدم وجود صوت	سنتز صوت خروجی مدل فردوسی	بدون نیاز به دیتاست	بدون نیاز به دیتاست
سعدی ❋❋	عدم وجود صوت	مدلسازی TTS	۱:پنج ساعت دیتاست متن ۲: پنج ساعت دیتاست از شعر به همراه صوت خوانش.	می توان از سایت گنجور برای ساخت دیتاست استفاده کرد
مولانا	عدم وجود صوت	سنتز خروجی مدل TTS سعدی	بدون نیاز به دیتاست	بدون نیاز به دیتاست
نظامی❋❋❋	عدم وجود صوت	مدلسازی TTS	۱:پنج ساعت دیتاست متن ۲: پنج ساعت دیتاست از شعر به همراه صوت خوانش.	می توان از سایت گنجور برای ساخت دیتاست استفاده کرد

[1] https://ganjoor.net/saadi/boostan/bab6/sh1#recitations

برای خوانش هر متن/شعر، چند صدا وجود دارد. برای هر شاعر می تون از یکی از این اصوات استفاده نمود.

پیرزن	کاراکتر عمومی	مدلسازی TTS	ساخت دیتاست	در ساخت دیتاست می توان از صدای کاراکترهای کارتونی استفاده کرد.
دختربچه	کاراکتر عمومی	سنتز خروجی صدای پیرزن	بدون نیاز به دیتاست	بدون نیاز به دیتاست
پیرمرد	کاراکتر عمومی	مدل گنجی	بدون نیاز به دیتاست	بدون نیاز به دیتاست
جادوگر****	کاراکتر عمومی	سنتز صوت/ مدلسازی TTS	درصورت سنتز صوت: بدون نیاز به دیتاست	درصورت سنتز صوت: بدون نیاز به دیتاست
پسربچه	کاراکتر عمومی	سنتز صوت یکی از مدل های فوق	بدون نیاز به دیتاست	بدون نیاز به دیتاست
غول	کاراکتر عمومی	سنتز صوت/ مدلسازی TTS	درصورت سنتز صوت: بدون نیاز به دیتاست	درصورت سنتز صوت: بدون نیاز به دیتاست

نکته: منظور از صوت با کیفیت، صوت بدون نویز، صدای پس زمینه، اکو، Distortion و صوت دیگران است. درصورتیکه این افکت ها در صوت وجود دارد، تنها در صورت حذف کامل آنها صوت با کیفیت بدست می آید. منظور از Distortion، مشکل نبودن فرکانس های ۶ هزار تا ۱۰هزار هرتز در صوت است. صوت باید شامل تمامی فرکانس ها از ۰ تا نزدیک ۱۰ هزار هرتز در نمایش spectrogram خود باشد. نیاز است تا صوت ها در Sample rate ۴۴۱۰۰هرتز و یا بیشتر و فرمت wav باشند.

*: درمورد صدای شهید مطهری پس از بررسی تمامی اصوات موجود از ایشان می توان روش دقیق را مشخص کرد. درصورتیکه حتی به اندازه ۲۰ دقیقه صدای با کیفیت از ایشان نباشد، می بایست از روش سنتز صوت و یا روش مدلسازی روی صدایی مشابه استفاده کنیم.

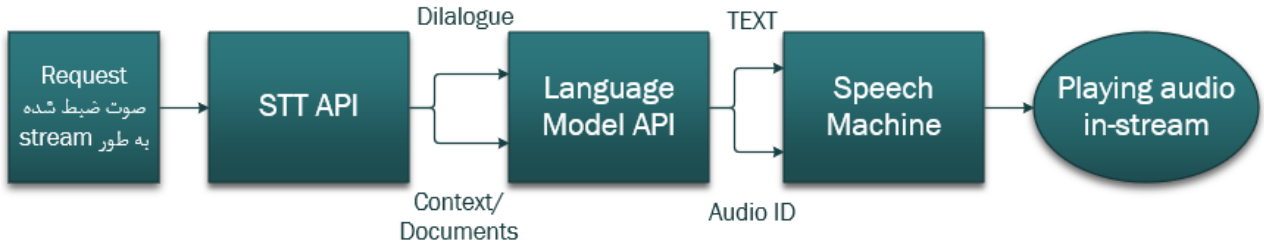
**: درمورد شعرا مانند سعدی، درصورتیکه هم نیاز به این باشد که امکان خواندن نثر و هم شعر را داشته باشند، برای کیفیت بهتر، نیاز است دو مدل، یکی برای خوانش نثر عادی و یکی برای خوانش شعر، ایجاد شود. برای ساخت این دو دیتاست می توان از سایت گنجور(لینک قرار گرفته در فوتر صفحه قبل) استفاده کرد.

***: درصورتیکه همه شعرا به طور همزمان در یک صحنه از بازی حضور ندارند و حضور حتی یکی از آنها به طور مجزا در صحنه دیگری است، می توان از مدل صوتی شعرای دیگر برای وی استفاده کرد و نیازی به مدلسازی مجدد نیست. بدیهی است که اطلاع از تعداد دقیق کاراکترها و همچنین حضور همزمان آنها در بازی برای این تصمیم گیری و تخمین لازم است که این اطلاعات در اختیار مجری پروژه قرار نگرفته است. این مورد برای کاراکترهای دیگر نیز صادق است.

****: برای کاراکتر جادوگر و غول، پس از تکمیل مدلسازی های مربوط به صداهای دیگر، ابتدا روش سنتز صوت روی مدل های آموزش داده شده مورد بررسی قرار خواهد گرفت و سپس درصورتیکه نیاز به کیفیت بالاتری بود، از صدای یکی از کاراکتر های داستانی مشابه این شخصیت ها، برای ساخت دیتاست و مدلسازی استفاده خواهد شد.

نکته۲: صدای مرد جوان و زن جوان، جز صداهای اعلامی برای این پروپوزال نبود اما دیتاست های آماده ای از این دو صدا برای مدلسازی موجود است.

نمودار یکپارچگی مدل های ساخت صوت(سنتز صوت و TTS)، مدل زبانی و مدل STT:



در نمودار قبل، منظور از audio id، کد شناسایی مربوط به هر یک از مدل های توسعه داده شده TTS، مانند مدل های موجود در جدول است.

۹- روش تحقیق (فرضیه ها و مبانی علمی و فنی پروژه)

این فرآیند شامل بررسی دانشی، جمع‌آوری اطلاعات و سپس پردازش داده‌های صوتی از نظر ویژگی های سنتز صوتی، جهت پیاده‌سازی سیستمی جهت ساخت گفتار انسانی است.

۱. فرض شده است که با استفاده از روش Voice Cloning می توان از یک دیتاستِ با کیفیت، دیتاستی جدید با صدای فردی خاص در زبانی جدید تولید کنیم که می تواند جهت ساخت مدل TTS با کیفیت مناسب استفاده شود.
۲. زمان اجرای ماژول سنتز صوت بر روی یک صوتِ ورودی، کمتر از 1s است. درواقع فرض شده است که طول صوت های ورودی و گفتار تولید شده کمتر از دقیقه باشد.
۳. در غیر این صورت فرض می شود زمان اجرای این ماژول(در مرحله Inference) اهمیتی ندارد(یعنی جهت استفاده Real-time نمی باشد.)
۴. فرض شده است که با تغییر فرکانس اصلی(Pitch) هر صوت و سرعت بیان آن می توان صوتی جدید با صدایی طبیعی تولید کرد.
۵. فرض شده است صوت هایی به عنوان ورودی(در نتیجهء مدل های TTS مختلف) در دسترس است تا سنتز صوت روی آنها اعمال شود.

۱۰-خروجی‌های حاصل از اجرای پروژه و چگونگی صحه‌گذاری بر آن‌ها (سنجه/شاخص):

- مدل های TTS آموزش داده شده
- فایل zip دیتاست های ساخته شده
- مدل Voice Cloning آموزش داده شده
- طراحی و توسعه سیستم مولد صوت گفتار انسانی
- سنجه: امکان تولید دو صوت متنوع، متفاوت و جدید از هر صوت ورودی، این صوت باید دارای کیفیتی طبیعی بوده و گفتاری مشابه انسان داشته باشد.
- مستندات پروژه: مراحل بررسی و پژوهش و نتایج آنها، مستندات فنی، نمونه های تست
- سورس کد ماژول سنتز صوت

منابع:

[1] Siedenburg K, Graves J, Pressnitzer D (2023) A unitary model of auditory frequency change perception. PLOS Computational Biology 19(1):e1010307. https://doi.org/10.1371/journal.pcbi.1010307

[2] Dobrowohl FA, Milne AJ, Dean RT. Controlling Perception Thresholds for Changing Timbres in Continuous Sounds. *Organised Sound*. 2019;24(1):71-84. doi:10.1017/S1355771819000074

۱۱-هزینه‌های پرسنلی:

ردیف	نام و نام‌خانوادگی	میزان تحصیلات			نوع همکاری							فعالیت محوله در گانت	ساعات همکاری در ماه	هرینه نفر ساعت (ریال)	هزینه نیروی انسانی در ماه (میلیون ریال)
		دانشجوی کارشناسی ارشد	کارشناسی ارشد	دکترا	رسمی		قراردادی								
					هیئت علمی	پژوهشگر	تمام وقت	ساعتی	مشاوره	حمایت از پایان‌نامه	طرح جایگزین/کسر خدمت				
۱	فاطمه وحیدیونسی		*			*		*				مجری تمامی مراحل گانت	۷۵	۲۵۰۰۰۰۰	۱۸۷۵۰۰۰۰۰
													۷۵	۲۵۰۰۰۰۰	۱۸۷۵۰۰۰۰۰
جمع کل هزینه‌های پرسنلی														۱۸۷۵۰۰۰۰۰ درماه مدت زمان: ۶ماه	

پیوست ۱

جدول زمانبندی (نمودار گانت)

عنوان پروژه: **ساخت سیستم سنتز گفتار انسانی** مدت پروژه(ماه): ۶ ماه

ردیف	کد فعالیت	شرح فعالیت	شرح فعالیت	فعالیت پیش نیاز	مدت به هفته	درصد وزنی	هزینه ها	ماه					
								۱	۲	۳	۴	۵	۶
۱	۱۰	بررسی و پژوهش	بررسی ویژگی های مختلف سیگنال صوتی	-	۱	۴							
			بررسی متد های مختلف سنتز صوت	۱۰-۱	۱	۴							
			بررسی مدل ها و دیتاست های فارسی موجود	-	۱	۴							
	۲۰	پیاده‌سازی و توسعه	ساخت ماژول و پیاده سازی متدهای مختلف سنتز صوت جهت ساخت گفتار انسانی	۱۰-۲	۱	۴							
			ساخت دیتاست های فارسی	۱۰-۳	۱۰	۴۰							
			آموزش مدل های Voice Cloning	۲۰-۲	۱	۴							
			مدلسازی TTS بر روی دیتاست های ساخته شده	۲۰-۲	۸	۳۲							
			بهینه سازی و بهبود زمان اجرا	۱و۴-۲۰	۱	۴							

۳	۳۰	یکپارچگی، تست و مستندسازی	تست ماژول و ساخت نمونه خروجی	۲۰-۴	۱	۴							
			مستندسازی	۳۰-۱									

❖ **اینجانب "فاطمه وحید یونسی" مدیر پروژه " ساخت سیستم سنتز گفتار انسانی" ، صحت مطالب مندرج در این شناسنامه را تأیید می نمایم.**

نام نام خانوادگی و امضاء مدیر پروژه:

فاطمه وحید یونسی