طرح سرویس تبدیل گفتار به نوشتار

شرکت فناوری اطلاعات و ارتباطات پاسارگاد آریان
ICT Holding of Pasargad Financial Group
www.fanap.ir

درخواست برای ارائه پیشنهاد طرح سرویس

«تبدیل گفتار به نوشتار»

مرکز مطالعات و تحقیقات شرکت فناپ

تاریخ تهیه سند: ۱۴۰۰/۰۵/۳۰

شماره ویرایش: ۱.۰

۱- مقدمه

هدف از این طرح تدوین مشخصات سرویس مورد نیاز شرکت فناوری اطلاعات و ارتباطات پاسارگاد آریان (فناپ) در حوزه تبدیل گفتار به نوشتار[1] بوده که می‌بایست توسط مجری یا مجریان احتمالی، طراحی و پیاده‌سازی گردد. در این راستا و در ادامه، خروجی طرح (Deliverables)، شرح نیازمندی‌ها و محدودیت زمانی اجرای طرح مشخص شده است. همچنین به مواردی که می‌بایست در تخمین بودجه مورد نیاز متقاضیان اجرای طرح لحاظ گردد، نیز اشاره شده است.

۲- مشروح مسئله تحقیقاتی

امروزه فناوری تبدیل گفتار به نوشتار خود را به عنوان راه حلی اساسی در خدمات نوشتاری معرفی کرده است. ارائه‌ی متن یک صوت با هزینه‌ی کم، آن هم به شیوه‌ای راحت و در عین حال دقیق و سریع جزء مهم‌ترین مزیت‌های این فناوری است. سرویس تبدیل گفتار به متن، تکنولوژیی است که اجازه می‌دهد انسان‌ها از صدای خود برای گفتگو با یک رابط کامپیوتری استفاده کنند. این موضوع به این دلیل اهمیت پیدا می‌کند که استفاده از ابزارهای دیجیتال مانند تلفن همراه، لپتاپ، رایانه‌ها و تبلت‌ها افراد را ملزم به تایپ کردن می‌کند و افراد روزانه زمان زیادی را صرف تایپ کردن در ابزارهای مختلف می‌کنند. به این ترتیب اگر نرم‌افزاری وجود داشته باشد که بتواند صوت یا گفتار (voice یا speech) را با دقت بالایی تبدیل به متن یا نوشتار (text) کند، مسلماً افراد زیادی از این نرم‌افزارها استفاده خواهند کرد و همین باعث صرفه جویی قابل توجهی در وقت آن‌ها خواهد شد. افراد بسیاری در سراسر دنیا هر روز از فناوری تبدیل و تشخیص صوت به متن استفاده می‌کنند. جستجوی صوتی گوگل و دستیارهای صوتی نظیر Siri و Alexa نمونه‌هایی از به‌کارگیری فناوری تبدیل صوت به متن هستند که جای خود را در زندگی روزمره‌ی انسان‌ها باز کرده‌اند.

کاربرد برنامه تبدیل صدا به متن (تشخیص گفتار) فقط منحصر به تایپ کردن گفتار یا صدا نیست. علاوه بر ایجاد یک نرم‌افزار تبدیل صوت به متن فارسی، می‌توان با استفاده از سرویس تبدیل صوت به متن (تبدیل گفتار به نوشتار)، هر نرم‌افزار یا سخت‌افزاری را که نیاز به تعامل با انسان و فهم گفتار و صحبت‌های افراد دارد توسعه داد و از این سرویس در کنار دیگر بخش‌ها استفاده نمود و به راحتی گفته‌های کاربر را به متن تبدیل کرد. استفاده از این سرویس، نیاز به تایپ کردن را از بین می‌برد و با سرعت و به راحتی یک متن تایپ شده و قابل ویرایش را در اختیار قرار می‌دهد.

فناوری تبدیل صوت به متن به دلیل فراهم نمودن امکان جستجو در بین فایل‌های صوتی و صداهای موجود در ویدیوهای موجود در شبکه‌های اجتماعی، کاربردهای مفیدی در این زمینه می‌تواند داشته باشد. تحلیل داده‌هایی که به این صورت از رسانه‌های دیجیتال دریافت می‌شود، منجر به دستیابی به دانش ارزشمندی می‌شود که به عنوان نمونه در برچسب‌گذاری فایل‌های صوتی و تصویری و پادکست‌ها و شناخت روندها و تحلیل نیازها و سلیقه‌های مشتریان کاربرد دارد.

۴- مسئله اصلی تحقیق و ضرورت آن

امروزه تولید محتوا در مشاغل از ارزش بالایی برخوردار است و کاربردهای مختلفی برای نویسندگان، روزنامه نگاران، وکلا، پزشکان، بخش روابط عمومی سازمان‌ها و ... دارد. نرم‌افزار تبدیل صوت به متن، ویژگی‌های خاصی مانند ویرایش، ذخیره فایل‌ها، امکان اشتراک‌گذاری، قابلیت دریافت فایل‌های صوتی در فرمت‌های مختلف، تبدیل آن به متن و همچنین اندازه‌گیری

[1] Speech to Text

دقت تبدیل فایل‌ها را نیز دارد. همچنین برنامه‌نویسان و توسعه‌دهندگان می‌توانند قابلیت تبدیل گفتار به متن را به وب‌سایت‌ها و اپلیکیشن‌های خود اضافه کنند.

یکی از مهم‌ترین کاربردهای این سرویس، استفاده در اپلیکیشن‌های مسیریاب برای خواندن آدرس‌ها و دستورالعمل‌های مسیریابی می‌باشد. به عنوان کاربردی دیگر، تبدیل گفتار به متن می‌تواند هم برای جستجوی صوتی و هم در محصولاتی نظیر اپلیکیشن‌های ترجمه مورد استفاده قرار گیرد. مشتریان این سرویس شامل بانک‌ها، اپراتورهای تلفن همراه، مراکز پشتیبانی، مراکز ارائه اطلاعات، مراکز درمانی، مراکز بیمه‌ای، مراکز رزرو بلیت و هتل، استارت‌آپ‌های حوزه خدمات و ... می‌باشند.

۵- تشریح نیازمندی‌های طرح

در این بخش ابتدا مشخصات اصلی طرح مطرح می‌شود و سپس روش انجام طرح، ورودی‌ها و خروجی‌های مورد انتظار و نحوه اعتبارسنجی طرح مشخص می‌گردد. در انتها نیز، قابلیت‌های فنی مورد نیاز در طرح تشریح می‌شود.

۵-۱- مشخصات اصلی

سرویس تبدیل صوت به متن برنامه‌ای است که با دریافت صدای صحبت انسان، آن را تفسیر کرده و تبدیل به متن می‌کند. در فناوری تبدیل صوت به متن، با دریافت فایل صوتی، صوت موجود در بازه‌های زمانی کوتاهی بررسی می‌شود و سپس با الگوریتم‌های موجود، مشخص شود هر صوت مربوط به چه حرفی است.

برای تبدیل فایل صوتی به متن از پردازش زبان طبیعی و یادگیری عمیق نیز استفاده می‌شود. از الگوریتم یادگیری عمیق به منظور تشخیص حروف و از پردازش زبان طبیعی به منظور بررسی خروجی و اصلاح آن در صورت نیاز استفاده می‌شود. برای مثال هم اکنون استفاده از مدل‌های زبانی[2] برای بررسی و اصلاح کلمات رواج زیادی دارد و این مدل‌ها توانایی قابل توجهی در اصلاح و یافتن شکل درست کلمات دارند.

مراحل مختلفی در روند تبدیل خودکار صوت به متن وجود دارد. هنگامی که انسان صحبت می‌کند، سیگنال‌های صوتی خارج شده از دهان او در مبدل آنالوگ لرزش ایجاد می‌کند. این لرزش‌ها توسط مبدل دریافت شده و به اطلاعات قابل فهم به زبان دیجیتال ترجمه می‌شوند. مبدل آنالوگ با انتخاب و اندازه‌گیری‌های مکرر و بسیار دقیق امواج صدا، یک فایل صوتی را تبدیل به داده‌های دیجیتالی می‌کند. این سیستم دارای یک فیلتر برای تشخیص صداهای مرتبط به صوت اصلی و تشخیص تغییر فرکانس‌ها است. همچنین قابلیت تنظیم سرعت گفتار و اصلاح صوت و همچنین تنظیم میزان صدا برای ارائه نتیجه‌ای بهتر و بهینه‌تر را دارد. مرحله بعدی شامل تقسیم سیگنال دریافتی به صدم یا هزارم ثانیه و تطبیق این قسمت‌های کوچک‌شده با الگوریتم اصلی ماشین است. سپس سیستم با استفاده از یادگیری ماشین، متن صوت را بر اساس آنچه قبلاً آموخته است ایجاد می‌کند. نتیجه می‌تواند به صورت یک فایل متنی قابل ویرایش ارائه شود.

۵-۲- ورودی سرویس

ورودی این سرویس دریافت صدا از طریق یک میکروفون یا یک فایل صوتی از پیش ضبط شده می‌باشد. فرمت‌های قابل قبول جهت ارسال صوت OGG، AAC، MP3، MP4، Wave، M4a می‌باشند.

[2] Language Model

۵-۳- خروجی مورد انتظار سرویس

این سامانه قادر خواهد بود اسناد متنی مستخرج از گفتار را ایجاد کند و پس از ویرایش آن توسط کاربر در قالب فایل‌های WORD و PDF در اختیار وی قرار دهد. سرویس تبدیل گفتار به نوشتار به صورت API یا REST Service باید عرضه شود. همچنین ارائه SDK اندروید و iOS به منظور بهره‌برداری آفلاین از نرم‌افزار در اولویت‌های بعدی مورد انتظار است.

۵-۴- نحوه اعتبارسنجی طرح

به منظور سنجش اعتبار طرح مورد نظر، دقت متن استخراج شده از صوت دریافتی سنجیده می‌شود. حداقل دقت مورد انتظار الگوریتم بکار رفته در سطح جمله های کوتاه و بر روی دیتاست هایی که به مجری طرح تحویل داده خواهد شد، ۹۸ درصد در سطح معیار SER و ارزیابی بر اساس معیار «تعداد جمله درست نسبت به تمامی جملات» است.

۵-۵- دیتاست‌های آموزشی

برخی از دیتاست‌هایی که برای آموزش الگوریتم می‌تواند مورد استفاده قرار بگیرد به شرح زیر است:

- Common Voice
- Farsdat
- LibriSpeech test-clean
- Libri-Light test-clean
- AISHELL-1
- DIRHA
- dev93

شایان ذکر است که دیتاست‌های معرفی‌شده صرفاً برای آموزش الگوریتم می‌توانند مورد استفاده قرار گیرند و دقت نهایی سامانه بر روی دیتاست تیم فنی فناپ ارزیابی خواهد شد.

۵-۶- قابلیت‌های مورد انتظار سرویس

نرم‌افزار تبدیل صوت به متن از فرمت‌های مختلف فایل‌های صوتی پشتیبانی کرده و کاربران می‌توانند صوت مورد نظر خود را در انواع فرمت‌های مختلف وارد این سرویس کرده و متن تبدیل شده و قابل ویرایش دریافت نمایند. سرویس تبدیل گفتار به متن با در اختیار داشتن داده‌های متنوع، درصد بسیاری از کلمات زبان فارسی را پوشش داده و تنوع واژگان مختلف زبان فارسی را شامل می‌شود. این سرویس همچنین بایستی دارای قابلیت‌های زیر باشد:

- ارائه سرویس به صورت آنلاین و وب سرویس، و نیز ارائه نرم‌افزار آفلاین در قالب اپلیکیشن اندروید و iOS
- قابلیت به‌کارگیری در نرم‌افزارها و پلتفرم‌های مختلف، و یا قابلیت استفاده به صورت مستقل و یا یکپارچه با سرویس‌های دیگر
- بهره‌مندی از دایره واژگان (فرهنگ لغت) بسیار وسیع
- پشتیبانی از انواع فرمت‌های صوتی
- قابلیت شخصی سازی دایره واژگان و سایر قابلیت‌ها متناسب با نیاز کاربر
- قابلیت تبدیل ورودی صوتی شماره‌های تماس به متن
- قابلیت به‌کارگیری علائم نگارشی برای درک بهتر جملات

- مستقل از گوینده و عدم نیاز به آموزش برای هر فرد
- قابلیت یادگیری لهجه و لحن بیان گوینده و شخصی‌سازی سرویس بر اساس صدا و بازه سنی و جنسیت و سایر مشخصه‌های صدای افراد؛ به این ترتیب، نرم‌افزار و سرویس گفتار به نوشتار هم می‌تواند بدون نیاز به train شدن به کار گرفته شود و هم می‌تواند بر اساس صدای کاربر نهایی شخصی‌سازی و train شود تا دقت آن افزایش یابد.
- ارائه API و SDK تبدیل گفتار به متن در قالب وب سرویس
- تبدیل گفتار به متن بصورت همزمان (Real-Time)
- تشخیص گفتار و صوت در محیط‌های نویزی
- پشتیبانی از انواع لهجه‌ها
- قابلیت تبدیل گفتار محاوره‌ای به متن
- پشتیبانی از انواع فرمت‌های صوتی و ویدیویی
- تبدیل گفتار انگلیسی به متن انگلیسی
- قابلیت تایپ صوتی اعداد به صورت عددی یا حروفی
- تایپ تمامی علائم نگارشی با صدا
- امکان افزودن کلمات جدید، انگلیسی یا عربی
- پشتیبانی از دستور گفتاری
- امکان پردازش برخط ورودی صدا (stream)
- گزارش دقت خروجی تولیدشده در هر واحد زمانی یا متنی، و نیز گزارش دقت خروجی کل

۶- گلوگاه‌های پروژه

یک محدودیت بزرگ در فناوری تشخیص خودکار گفتار، توانایی تولید متن به صورت کلمه به کلمه است. در غیاب هوش انسانی، سیستم تنها قادر به رونویسی آنچه که می‌شنود است و این بدان معناست که متنی که در نهایت سیستم تحویل می‌دهد ممکن است به‌هم‌ریخته و غیرقابل‌فهم باشد. مکث کردن هنگام مکالمه، ایجاد صداهایی که معنای خاصی نداشته و صرفاً حس را بیان می‌کنند و لغزش بر روی برخی از کلمات گفته شده بسیار معمول است. متن تولید شده توسط نرم‌افزار تبدیل گفتار، کلمه به کلمه و شامل تمام شنیده‌هایش خواهد بود.

مهم‌ترین جنبه‌ی منفی نرم‌افزار تبدیل گفتار به متن، دقت آن است. اگر محتوای پیچیده در فایل صوتی وجود داشته باشد، سیستم تشخیص خودکار گفتار ممکن است نتایج نامفهومی را ایجاد کند. برای دریافت خروجی با دقت از سرویس تبدیل گفتار به نوشتار، به فایل‌های صوتی تمیزی نیاز است. همچنین سیستم‌های تشخیص خودکار گفتار ممکن است صرفاً با یک زبان به صورت تخصصی کار کنند و اینجاست که نیاز به چند زبان خود را نمایان می‌کند. برخی از سیستم‌های تشخیص گفتار نیز ممکن است در شناسایی نام‌های تجاری و اصطلاحات خاص حوزه صنعت با مشکل مواجه شوند.

- **عام بودن گفتار به نوشتار**

یکی از مشکلاتی که کاربران فارسی زبان در استفاده از تایپ صوتی گوگل با آن مواجه هستند، آن است که سرویس گوگل تنها قادر است گفتار عام در زبان فارسی را متوجه شود و کلمات تخصصی در زبان فارسی که برخی کسب‌وکارها، همانند

وکلا و حقوقدانان و...، از آن‌ها به کرات استفاده می‌کنند را به درستی متوجه نمی‌شود. به همین خاطر کسب‌وکارهای این چنینی برای تبدیل گفتار به نوشتار و تایپ صوتی نمی‌توانند بر کمک گوگل اتکا کنند.

- **عدم درک لهجه‌های مختلف زبان فارسی**

زبان فارسی پر است از لهجه‌ها و گویش‌های مختلف. اگر سرویسی می‌خواهد در زبان فارسی به خوبی کار کند نیاز دارد که تمام این لهجه‌ها را متوجه شود. همان طور که در مورد قبل به آن اشاره شد تایپ صوتی گوگل گفتار عام زبان فارسی را متوجه می‌شود و باز هم کاربران فارسی زبان با مشکلات بسیاری در این زمینه روبه‌رو هستند.

۷- الزامات طرح

برنامه تایپ گفتاری فارسی، صحبت فارسی را به متن تبدیل می‌کند. تبدیل صوت به متن علاوه بر سرویس Web API و Server-side SDK در اولویت نخست، به صورت بسته نرم‌افزاری اندروید و iOS (Client-side SDK) نیز باید ارائه شود تا توسعه‌دهندگان و برنامه‌نویسان در سیستم‌عامل‌های مختلف بتوانند قابلیت تایپ گفتاری و فرمان صوتی را به نرم‌افزارهای خود اضافه نمایند. برای افزودن قابلیت تشخیص صدا به نرم‌افزار یا یک دستگاه، باید بتوان به عنوان ماژول یا سرویس از آن استفاده نمود.

- **تشخیص لهجه و گویش**

زبان فارسی پر است از گویش‌ها و لهجه‌های متفاوت. برای آنکه یک نرم‌افزار تبدیل گفتار به نوشتار بتواند به خوبی در زبان فارسی کار کند، باید بتواند تمام این گویش‌ها و لهجه‌ها را تشخیص دهد.

- **تشخیص گفتار رسمی و محاوره‌ای**

در زبان فارسی تفاوت بین گفتار محاوره و گفتار رسمی بسیار زیاد است و سرویس تبدیل گفتار به متن کاربردی باید بتواند این دو را از یکدیگر تمیز دهد.

- **سفارشی سازی تبدیل گفتار به نوشتار**

قابلیت سفارشی‌سازی سرویس برای کسب‌وکارهای مختلف باید وجود داشته باشد. برخی از کسب‌وکارها همانند حوزه وکالت دارای اصطلاحات تخصصی در زبان فارسی هستند که این کلمات تخصصی به سرویس تایپ صوتی گوگل آموزش داده نشده است.

- **قابلیت پردازش فایل‌های صوتی از پیش ضبط شده**

باید بتوان فایل‌های صوتی با فرمت‌های مختلف را به محصول داد و در مقابل متن این فایل‌ها را دریافت نمود.

- **قابلیت فرمان‌پذیری در خصوص علائم نگارشی**

سرویس تایپ گفتاری فارسی باید دارای استانداردهایی برای دریافت فرامین مربوط به درج انواع علائم نگارشی (مانند ویرگول، نقطه، دو نقطه، نقطه ویرگول، بولت، شماره‌گذاری، پرانتز و سایر علائم متداول) باشد

- **قابلیت تبدیل گفتار به نوشتار به صورت آفلاین**

محصول بایستی دارای اپلیکیشن اندرویدی و نسخه iOS باشد که کاربران مختلف بتوانند بدون نیاز به اتصال به اینترنت و به صورت آفلاین گفتار خود را به نوشتار تبدیل کنند.

۸- محدودیت زمانی اجرای طرح

این طرح پس از توافق طرفین می‌بایست حداکثر در یک بازه‌ی زمانی شش ماهه به اتمام رسد.

۹- تخمین بودجه اجرای طرح

تخمین بودجه اجرای طرح در قالب هزینه‌های نفر/ساعت نیروی انسانی، سربار و تجهیزات جانبی مورد نیاز می‌بایست مشخص شود.