ارائه راهکاری عملی جهت کاهش خستگی ذهنی ناشی از شنیدن گفتار ماشینی یکنواخت

حمید رضا نیرومند[1]

چکیده

با پیشرفت تکنولوژی، سیستم‌های تبدیل متن به گفتار هر روز بیشتر و بیشتر مورد استقبال قرار می‌گیرند، این سیستم‌ها به منظور قرائت متون برای استفاده در رایانه‌ها طراحی شده‌اند و هر روز بیشتر به استفاده از آن در دیگر سیستم‌های هوشمند توجه می‌شود. تبدیل متن به گفتار با دو قالب اساسی زبان یعنی قالب نوشتاری و قالب گفتاری آن در ارتباط است. در قالب نوشتاری با پیچیدگی‌های فراوان در انواع تحلیل‌های صرفی، نحوی، معنایی به منظور استخراج انواع اطلاعات مورد نیاز از متن ورودی رو به رو هستیم و قالب گفتاری نیز با هدف دستیابی به گفتاری نزدیک به گفتار طبیعی، با مشکلات عدیده‌ای در تولید تلفظ و نوای مناسب برای گفتار بازسازی شده رو به رو می‌باشیم.

اما اگر مشکلات ذکر شده را تماماً حل شده بدانیم، یکی دیگر از چالش‌ها در تولید نرم افزارهای مبدل متن به صدا، این است که صدای تولیدی به طور خسته کننده‌ای، یکنواخت است. به عبارت دیگر حجم صدای تولید شده، از ابتدا تا انتهای متن، در یک سطح بوده و این امر در متون طولانی بسیار خسته کننده خواهد بود. در این مقاله، بر آنیم تا یک راهکار عملی برای کاهش اثر خستگی در صدای تولید شده توسط الگوریتم‌ها و نرم افزارهای تبدیل متن به صدا ارائه دهیم.

واژه‌های کلیدی

تبدیل متن به گفتار، بهبود صدای یکنواخت ماشینی، گفتار ماشینی، نرم افزار متن‌خوان، تنوع مصنوعی در صدا

Suggesting a method for solving monotone voice problem in TTS

Hamid Reza Niroomand

Abstract

One of the most attractive research subjects nowadays is “Text To Speech”. People’s interest in “Siri” on Apple iPhone showed that people love talking and listening to machine! But the biggest problem in this case is that the voice produced by a machine is monotone and boring. There are dozens of articles about prosody in TTS but the main problem is that even if we consider this approach has been implemented as good as a real human voice, but we still have monotone voice problem even in between humans! Many audiences claim the speakers about their monotone voice.

This research is something beyond the prosody of voice. It tries to suggest a solution to decrease the mental fatigue while listening to a monotone sound. In one sentence it suggests developers to create an artificial diversity on sounds.

Keywords

Monotone Speech Problem, Text To Speech, TTS, artificial diversity on sound.

[1] مربی، دانشگاه آزاد اسلامی واحد زرند، info@niroomand.ir

۱- مقدمه

با پیشرفت تکنولوژی، سیستم‌های تبدیل متن به گفتار هر روز بیشتر و بیشتر مورد استقبال قرار می‌گیرند، این سیستم‌ها به منظور قرائت متون برای استفاده در رایانه‌ها طراحی شده‌اند و هر روز بیشتر به استفاده از آن در دیگر سیستم‌های هوشمند توجه می‌شود. پیش‌بینی می‌شود به زودی در اکثر وسایل، قطعات هوشمندی تعبیه شود که از سیستم پردازش صوت و همچنین تبدیل متن به گفتار برای برقراری ارتباط با انسان استفاده می‌کنند. برخی از کاربردهای سیستم‌های تبدیل متن به گفتار[1] عبارتند از:

- کاربرد در بازی‌ها
- کاربرد در تلفن‌های گویا (اعلام وضعیت مسیر، آب و هوا، موجودی حساب بانکی و غیره)
- خواندن و صحبت کردن برای معلولین
- آموزش‌دهنده گویا
- و ده‌ها کاربرد دیگر

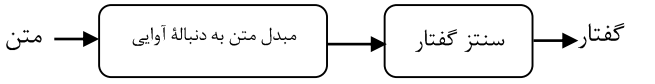
برای تولید گفتار در TTS بایستی از روش‌ها و الگوریتم‌هایی جهت خواندن متون استفاده شود چرا که ذخیره‌سازی تمامی کلمات یک زبان (با توجه به تعداد بسیار زیاد کلمات و رشد همیشگی آن) غیرممکن و در راستای تولید گفتار طبیعی بی‌فایده خواهد بود. «زیر و بمی» یا تغییرات «فرکانس گام»[2]، «دیرش»[3] ، «شدت»[4] و نیز «درنگ»[5] چهار عنصر نوایی گفتار هستند که معمولاً در سطوح مختلف اعم از هجا، واژه یا جمله اثر خود را نشان می‌دهند و در سیستم‌های TTS بایستی گنجانده شوند.

اساساً تمامی سیستم‌های تبدیل متن به گفتار دارای دو بخش اصلی هستند که عبارتند از [1]:

۱- استخراج اطلاعات آوایی و سایر اطلاعات زبانی مانند تکیه و نوا از متن ورودی. (مبدل متن به دنبالهٔ آوایی)

۲- تبدیل این اطلاعات آوایی به شکل موج گفتار. (سنتز گفتار)

که نحوه ارتباط این دو بخش و تولید صـــدا از روی یک متن ورودی به صورت کلی در شکل زیر نشان داده شده است.



شکل1 - شکل کلی ارتباط بخش‌های مختلف یک سیستم تبدیل متن به گفتار

دو رویکرد رایج برای تولید گفتار وجود دارد، اولی روش سنتز مبتنی بر قاعده است که در آن پارامترهای مشخصه گفتار در هر بازهٔ زمانی توسط مجموعه‌ای از قواعد تولید می‌شوند و بعدی روش اتصال قطعات گفتار که در آن واحدهای از پیش ذخیره شده‌ی صوتی برای تولید عبارتی دلخواه در کنار هم چیده می‌شوند. نمونهٔ مشهور سنتز کننده‌های مبتنی بر قاعده، Klatt و مدل تجاری‌تر آن DECTalk می‌باشد.

۲- تعریف تبدیل متن به گفتار

تولید گفتار به معنای تولید مصنوعی گفتار انسان است. یک سیستم تبدیل متن به گفتار یک سیستم مبتنی بر سخت‌افزارهای دیجیتال یا رایانه است که باید توانایی خواندن متن‌های یک یا چندین زبان را داشته باشد. این متون ممکن است از طریق صفحه کلید و یا به صورت یک فایل متنی و یا پس از شناسایی توسط یک سیستم شناسایی نوری نویسه‌ها[6] دریافت شوند. در حقیقت تبدیل متن به گفتار تلاش برای تقلید توانایی‌های انسان در خواندن متون است. [2]

گفتار می‌تواند از اتصال قطعات گفتاری ضبط شده که در یک دادگان ذخیره شده‌اند و یا از نتیجه اعمال یک سیگنال تحریک مناسب به فیلتری که پارامترهای آن از واحدهای گفتار طبیعی به دست آمده‌اند تولید شود. سیستم‌های تولید گفتار از جهت نوع واحد گفتاری ذخیره شده و یا روش بکار رفته در استخراج پارامترهای گفتار و نحوه استفاده مجدد از آن‌ها برای تولید گفتار متفاوت هستند.

واحدهای گفتاری واج یا دوآوایی قادر به تولید محدوده وسیعی از گفتار هستند. لیکن دسترسی به کیفیت بالایی از طبیعی بودن گفتار تولیدشده مشکل است. در کاربردهای خاص، ذخیره مستقیم کلیه کلمات و حتی جملات و استفاده مجدد از آن‌ها کیفیت بالایی را در تولید گفتار باعث می‌شود. اما این روش با محدودیت تعمیم‌پذیری در تولید متونی با محتوایی متفاوت با آن‌چه که کلمات و جملات آن را ضبط نموده‌ایم مواجه است.

کیفیت خروجی یک سیستم تبدیل متن به گفتار معمولاً با میزان شباهت آن با صدای انسان و میزان قابلیت فهم آن سنجیده می‌شود. خروجی یک سیستم تبدیل متن به گفتار با وضوح کافی می‌تواند علاوه بر نیاز افراد نابینا، لال و افراد با ناتوانی خواندن، کاربردهای عمومی گسترده‌ای داشته باشد.

۳- چالش‌ها بر سر راه تبدیل متن به گفتار

تبدیل متن به گفتار با دو قالب اساسی زبان یعنی قالب نوشتاری و قالب گفتاری آن در ارتباط است. در قالب نوشتاری با پیچیدگی‌های فراوان در انواع تحلیل‌های صرفی، نحوی، معنایی به منظور استخراج انواع اطلاعات مورد نیاز از متن ورودی رو به رو هستیم. بسیاری از این تحلیل‌ها در کاربردهای دیگری چون ترجمه ماشینی، خلاصه‌سازی و طبقه‌بندی متون، سیستم‌های پرسش و پاسخ، بازیابی اطلاعات، جستجو در پایگاه‌های داده متنی و مانند آن نیز مورد نیاز است. [3]

[1] Speech Synthesis / Text-to-Speech (TTS)

[2] Pitch

[3] Duration

[4] Intensity

[5] Pause

[6] Optical Character Recognition (OCR)

علاوه بر این در تبدیل متن به گفتار با قالب گفتاری زبان نیز سر و کار داریم و لذا با هدف دستیابی به گفتاری نزدیک به گفتار طبیعی، با مشکلات عدیده‌ای در تولید تلفظ و نوای مناسب برای گفتار بازسازی شده رو به رو می‌باشیم. این مشکلات همواره مورد توجه محققان در سراسر دنیا قرار داشته است و مقالات فراوانی همه ساله در بسیاری از کنفرانس‌های علمی معتبر ارائه می‌گردد. با وجود تحقیقات فراوان و سرمایه‌گذاری‌های بسیار در این زمینه، پیچیدگی ذاتی مسائل در این حوزه و گستردگی علوم و فناوری‌های مرتبط با آن سبب شده است که راه زیادی تا تولید محصولات ایده آل تبدیل متن به گفتار باقی باشد.

در این میان، یکی دیگر از چالش‌ها در تولید نرم افزارهای مبدل متن به صدا که هنوز فکری به حال آن نشده است، این است که صدای تولیدی به طور خسته کننده‌ای، یکنواخت است. به عبارت دیگر حجم صدای تولید شده، از ابتدا تا انتهای متن، در یک سطح خواهد بود و این امر در متون طولانی (به طور مثال، اگر از نرم افزار خواسته شود که یک مقاله یا یک داستان را بخواند) بسیار خسته کننده خواهد بود.

ما فکر می‌کنیم یکی از دلایلی که باعث می‌شود ذهن مخاطب از شنیدن صدای طبیعی انسان خسته نشود اما از شنیدن صدای ماشینی احساس خستگیِ ذهنی کند، این است که هنگامی که یک انسان یک مقاله یا داستان را می‌خواند، به طور هوشمندانه حجم صدای خود را کم و زیاد می‌کند و یا گاهی عواملی مانند جا به جا شدن شخص و یا چرخشِ سرِ او، باعث این کم و زیاد شدن حجم صدا می‌شود و بنابراین تنوع در حجم صدا، مانع خستگی ذهنی مخاطب می‌شود.

در این تحقیق، بر آنیم تا بر اســاس همین پدیده طبیعی، یک راه حل مصــنوعی برای کاهش اثر خســتگی در صــدای تولید شــده توسط الگوریتم‌ها و نرم افزارهای تبدیل متن به صدا ارائه دهیم.

۴- اهمیت و ضرورت پژوهش در این زمینه

گفتار، ابزار اولیه ارتباط بین انسان‌هاست. انسان‌ها به کمک گفتار خود می‌توانند مفاهیم متفاوتی را به مخاطبان خود انتقال دهند. گفتار علاوه بر این که ابزاری مناسب برای انتقال دانسته‌ها و بیان نیازها است، بهترین روش انتقال مفاهیم ذهنی و احساسات درونی نیز به حساب می‌آید. در واقع برتری بارز گفتار بر نوشتار متناظر با آن این است که گفتار اطلاعات جانبی بیشتری را به شنونده انتقال می‌دهد که بعضاً هدف اصلی بیان جملات نیز انتقال همین اطلاعات می‌باشد.

یکی دیگر از ویژگی‌های بارز گفتار سهولت انتقال آن به مخاطب است. به این ترتیب که در مواقعی مناسب‌تر است ما به جای انتقال تصویری یا نوشتاری اطلاعات آن‌ها را به صورت گفتاری منتقل نماییم تا بازه وسیع‌تری از مخاطبان را در برگیرد. برای مثال در اماکن عمومی که نصب تابلوهای اطلاعاتی به شکل گسترده امکان‌پذیر نمی‌باشد، انتقال اطلاعات از طریق بلندگوها یک گزینه کم هزینه است. از طرف دیگر این نوع انتقال پیام موجب می‌شود که مخاطب زودتر در جریان قرار گیرد. [4]

قابلیت‌های فوق الذکر باعث جلب توجه و علاقه بسیاری به سمت ایجاد سیستمی جهت تولید گفتار شده است. تمام این مطالب در کنار کاربرد وسیع سیستم‌های تولید گفتار در عرصه زندگی انسان، باعث رشد روز افزون زمینه‌ای از مباحث پردازش‌های هوشمند تحت عنوان تبدیل متن به گفتار گردیده است.

از طرفی، بر اساس آنچه در مطالب آینده خواهد آمد، انسان از صدای ناهنجار دور می‌شود و به صداهای خوشایند نزدیک. همه این‌ها تأثیر صدا در زندگی انسان و جذب شدن و یا نشدن او به یک موضوع خاص را می‌رساند. حال، چطور می‌توان انتظار داشت کاربران به سمت استفاده از مبدل‌های متن به گفتار جذب شوند در حالی که با شنیدن صدای تولید شده توسط آن‌ها احساس خستگی می‌کنند؟

در این مقاله ما بر آنیم که تا راه حلی ارائه کنیم که تا حد ممکن از این اثر منفی بکاهیم و در نتیجه‌ی آن، رغبت کاربران برای استفاده از نرم افزارهای تبدیل متن به گفتار افزایش یابد.

۵- تأثیرات صدا بر روی مغز انسان [5]

صدا تمام مدت روی ترشح هورمون‌های بدن انسان، همچنین نفس کشیدن، ضربان قلب و امواج مغزی اثر می‌گذارد. دانشمندان معتقدند صدا به چهار صورتِ اساسی روی انسان اثر می‌گذارد:

۱- به صورت فیزیولوژیکی: به طور مثال، با شنیدن صدای زنگ ساعت بلند یک افزایش کورتیزول[1]، همان هورمون ستیز / گریز، در انسان ایجاد می‌شود. فقط صداهای ناهنجار نیستند که این اثر را دارند. به طور مثال، بسامد صدای امواج دریا تقریباً ۱۲ چرخه در دقیقه است. برای خیلی‌ها این صدا، آرامش‌بخش است، و جالب است که ۱۲ چرخه در دقیقه تقریباً فرکانس نفس کشیدن انسان در زمان خوابیدن است. یعنی انسان احساس می‌کند که در حال استراحت است و استرسی ندارد، و یا در تعطیلات به سر می‌برد.

۲- به صورت روانی: موسیقی قوی‌ترین نوع از صدا است که روی احساسات انسان اثر می‌گذارد. به طور مثال برخی موسیقی‌ها انسان را غمگین می‌کنند و برخی شاد. البته موسیقی تنها صدایی نیست که روی احساسات انسان اثر می‌گذارد. صداهای طبیعی هم همین اثر را دارند. برای مثال آواز پرندگان برای بیشتر مردم اطمینان‌بخش است. چرا که در طول صدها هزار سال، انسان یاد گرفته است که زمانی که پرندگان می‌خوانند، همه چیز امن است و زمانی که نمی‌خوانند باید نگران شد.

۳- به صورت ادراکی: همه ما متوجه صحبت‌های دو نفر که به طور هم زمان صحبت می‌کنند نمی‌شویم. انسان برای پردازش ورودی

[1] cortisol

صدا، پهنای باند خیلی کمی دارد برای همین صدایی مثل سر و صدا در دفتر کار راندمان او را به شدت کاهش می‌دهد.

۴- به صورت رفتاری: با توجه به آن همه تأثیراتی که در بالا بیان شد، اگر رفتار انسان تغییر نکند، چیز عجیبی خواهد بود. برای مثال می‌توان به راحتی حدس زد که راننده‌ای که در ماشین خود آهنگ تکنو گوش می‌کند، بعید است که با سرعت عادی ۴۰ کیلومتر در ساعت رانندگی کند!

در مجموع، انسان از صدای ناهنجار دور می‌شود و به صداهای خوشایند نزدیک و این تنها چیزی نیست که صداهای ناهنجار به آن آسیب می‌رساند! بر اساس تحقیقات، بیشتر صداهای فروشگاه‌های خرده فروشی، نامناسب، تصادفی و حتی خصومت آمیز محسوب می‌شوند و اثر باور نکردنی روی فروش دارند. برخی فروشگاه‌ها ۳۰ درصد از فروش خود را با خارج شدن سریع مشتری‌ها از مغازه یا حتی برگشتن آن‌ها از مقابل در فروشگاه، از دست می‌دهند.

۶- پیش‌زمینه و راهکارهای مرتبط

آنچه از بررسی ده‌ها مقاله فارسی و انگلیسی بر می‌آید این است که بسیاری از محققان در زمینه تبدیل متن به گفتار به این مشکل پی برده‌اند که صدای تبدیل شده با الگوریتم‌های تبدیل متن به گفتار، برای مخاطب، کمی نامأنوس و خسته کننده است. اما تمامی آن‌ها به اتفاق، به این نتیجه رسیده‌اند که اگر به صدای تبدیل شده، لحن نیز اضافه گردد، این خستگی ذهنی کاهش می‌یابد. این در حالی است که اولاً به ازعان بسیاری از محققان، پیاده‌سازی لحن در TTS بسیار پیچیده است و ثانیاً ما در سخنرانی‌هایی که یک انسان به طور طبیعی صحبت می‌کند نیز شاهد هستیم که با اینکه او جملاتش دارای لحن است اما اگر به طور یکنواخت صحبت کند، مخاطبان خیلی سریع خسته خواهند شد. با توجه به تعریف صدای یکنواخت، می‌توان به این نتیجه رسید که حتی اگر لحن نیز در یک مبدل متن به گفتار پیاده‌سازی شود، باز هم نیاز به تغییر حجم صدا لازم به نظر می‌رسد.

۶-۱- تعریف و تأثیر «لحن» در روان‌شناسی گفتار [6]:

لحن، همان جنبه موسیقایی از صدای انسان است که به بمی، حجم، سرعت و تأکید اشاره دارد. مخاطبان به طور غریزی به لحن صدا واکنش نشان می‌دهند (گاهی واکنش مثبت و گاهی منفی).

تحقیقاتی که توسط دپارتمان روان‌شناسی دانشگاه پیتس‌برگ در آمریکا[1] انجام شده است نشان می‌دهد که مردم به طور غریزی بر اساس لحن صدای صحبت کننده در مورد او قضاوت می‌کنند. به خصوص، انسان‌ها فکر می‌کنند کسانی که صداهای عمیق‌تری دارند برای کارهای مدیریتی بهتر هستند و برعکس، صدای نازک و تیز در دیگران حس ناتوانی در مدیریت ایجاد می‌کند.

صدای یکنواخت به صدایی گفته می‌شود که از ابتدا تا انتها در یک سطح باقی می‌ماند و هیچ تنوعی در فشار در آن وجود ندارد. روان‌شناسان سخنرانی معتقدند: با تغییر دادن تُن صدا، زندگی و انرژی به پیغامی که قصد انتقال آن‌را دارید اضافه می‌شود. احساساتی مانند شور و اشتیاق و طنز و هیجان می‌تواند از لحن صدای گوینده به مخاطب منتقل شود، بنابراین می‌توان انتظار داشت که صدایی که دارای لحن درستی نیست، روی مخاطب تأثیرات منفی مختلفی خواهد گذاشت و این، همان چیزی است که باعث می‌شود خیلی از کاربران به استفاده از نرم افزارهای تبدیل متن به صدا رغبت نداشته باشند.

۶-۲- لحن در TTS

بخش مهمی از اطلاعات صوتی موجود در گفتار به نوای گفتار بر می‌گردد و نوای گفتار علاوه بر اینکه حاوی اطلاعات مهم گفتاری اطراف جهت تأکیدات و نوع جمله از جهت سوالی، خبری بودن است، در طبیعی بودن گفتار نیز تأثیر به‌سزایی دارد. در زمینه زیر و بمی و نوای گفتار فارسی، پژوهش‌های متعددی انجام شده است که در [2] در صفحه ۳۳۹ لیست کاملی از آن‌ها بیان شده است. اما با وجود اینکه هنوز کار عملی قابل قبولی در این زمینه انجام نشده است، حتی اگر فرض کنیم لحن در تبدیل متن به گفتار به خوبی پیاده‌سازی شده است، ضرورت این پژوهش همچنان بر جای خود باقی‌ست.

۷- توصیف راهکار پیشنهادی

۷-۱- ایده اول: کاهش و افزایش صدا به صورت سینوسی

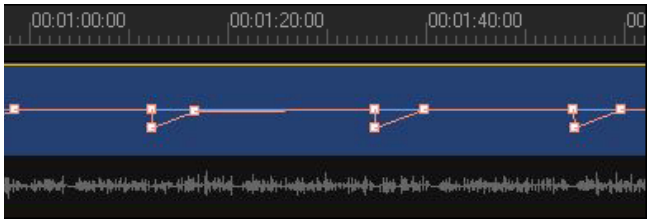
همانطور که پیش از این توضیح داده شد، در راهکار ارائه شده در این تحقیق، ما به دنبال ایجاد تنوع مصنوعی بر روی گفتار یکنواخت هستیم. برای انجام این کار، اولین ایده‌ای که به ذهنمان رسید این بود که حجم صدا را بدون توجه به کلمات و جملات بیان شده، به صورت سینوسی کم و زیاد کنیم. اما پس از انجام آزمایشات مختلف، متوجه شدیم که در صورت انجام چنین کاری، شنونده کاملاً متوجه کاهش و افزایش صدا خواهد شد و همین موضوع باعث می‌شود حواس کاربر از توجه به متن، به افزایش و کاهش حجم صدا معطوف شده و این مشکل به نامفهوم بودن طبیعی صداهای تبدیل شده از متن به گفتار اضافه شود و در نتیجه شنونده به سختی متوجه مفهوم متن خواهد شد.

۷-۲- ایده دوم: کاهش در سکوت، افزایش در ۵ ثانیه

بنابراین در مرحله دوم، ایده جدیدی را به کار گرفتیم: با توجه به حافظه موقت شنونده، صدا را در سکوت به اندازه ۲۰ درصد کاهش دهیم و در

[1] Department of Psychology, University of Pittsburgh , USA

عرض ۵ ثانیه به حالت اول برگردانیم. منظور از سکوت در تبدیل متن به گفتار می‌تواند لحظه رسیدن به یک نقطه و یا ویرگول و یا هر پایان‌دهنده یا ایجاد کننده‌ی مکث دیگری باشد و در یک سخنرانی یکنواخت، سکوت می‌تواند یک لحظه وقفه (هر چند بسیار کوتاه) در صحبت‌های سخنران باشد. در این صورت شنونده در مدتی که سکوت رخ داده، حجم صدای قبلی را تا حد زیادی فراموش خواهد کرد و از لحظه‌ی آغازِ مجددِ صدا، بدون اینکه حواس او پرت شود، صدا با حجم کمتری پخش خواهد شد و به مرور و در عرض ۵ ثانیه صدا به حالت ۱۰۰ درصد خواهد رسید. تصویر زیر، پیاده‌سازی این ایده را بر روی یک فایل صوتی نمونه نشان می‌دهد:



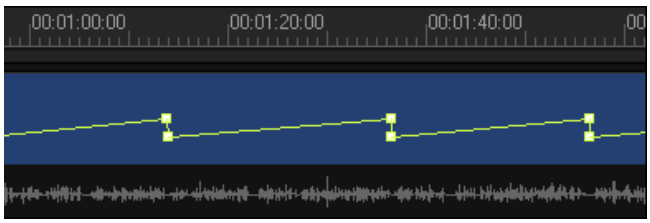
شکل ۲ - ایده‌ی دوم کاهش در سکوت و افزایش به نرمی

این تغییرات در حجم می‌تواند به صورت متناوب، در هر ۲۰ تا ۶۰ ثانیه یک بار اتفاق بیفتد. البته اگر در این فاصله یک سکوت یافت شود، در غیر این صورت یک دوره ۶۰ ثانیه‌ای دیگر به دنبال سکوت گشته می‌شود.

نتیجه‌ی این ایده هر چند بهتر از ایده‌ی اول بود اما با توجه به کوتاه بودن مدت اوج‌گیری صدا (۵ ثانیه) مجدداً این کاهش و افزایش محسوس بود، بنابراین این امکان وجود داشت که حواس شنونده متوجه تغییرات حجم صدا شود.

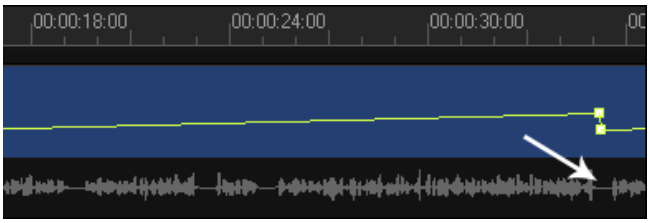
۷-۳- ایده سوم: کاهش در سکوت، افزایش تا سکوت بعدی

در مرحله سوم، ایده‌ی دوم کمی تغییر کرد: همانند مرحله دوم، حجم صدا در سکوتی که در فاصله زمانی ۲۰ تا ۶۰ ثانیه قرار دارد به اندازه ۲۰ درصد کاهش خواهد یافت اما اوج‌گیری صدا با طول زمانی بیشتری اتفاق خواهد افتاد. به طور دقیق، صدا تا سکوت بعدی و به مرور افزایش خواهد یافت. تصویر زیر، پیاده‌سازی این ایده را بر روی یک فایل صوتی نمونه نشان می‌دهد:



شکل ۳ - ایده‌ی سوم کاهش حجم صدا در سکوت و افزایش به نرمی

در تصویر زیر که بزرگنمایی شده‌ی تصویر بالا است، کاهش حجم صدا در سکوت، بهتر مشخص است:



شکل ۴ - بزرگ‌نمایی ایده‌ی سوم

پس از انجام آزمایش بر روی چند نمونه محدود به این نتیجه رسیدیم که این همان چیزی است که مد نظر این تحقیق است؛ یعنی تنوع در صدا به طوری که مخاطب در عین حال که از این تنوع راضی بوده و احساس آرامش می‌کند، حواس او متوجه این تنوع در حجم صدا نباشد.

۷-۴- یافتن سکوت در کاربرد تبدیل متن به گفتار

یافتن سکوت در نرم افزارهای تبدیل متن به گفتار ساده‌تر از یافتن سکوت در فایل سخنرانی یکنواخت خواهد بود. در تبدیل متن به گفتار می‌توان مدت زمان تولید و یا پخش فایل صوتی مربوط به یک صامت و مصوت را به طور میانگین محاسبه کرد، سپس ۲۰ ثانیه را تقسیم بر مدت زمان به دست آمده کرد تا تعداد صامت و مصوت‌هایی که باید پخش شود و سپس به دنبال سکوت گشته شود به دست آید. به طور مثال اگر طول فایل صوتی مربوط به هر صامت و مصوت را یک بیست و پنجم ثانیه بگیریم، باید ۵۰۰ ترکیب صامت و مصوت بگذرد و سپس شروع به جستجوی سکوت (نقطه، ویرگول، علامت تعجب و غیره) کنیم.

۷-۵- یافتن سکوت در کاربرد صدای آنالوگ ضبط شده

یافتن سکوت در فایل‌های ضبط شده از حالت آنالوگ یکی از پیچیده‌ترین مباحث در علم کامپیوتر است و نیاز به محاسبات ریاضی بسیار پیچیده‌ای دارد، چرا که ابتدا باید تعریف مشخصی از «سکوت» ارائه شود و این تعریف باید نویزها و صدای پس‌زمینه را مد نظر قرار دهد. اما با توجه به اینکه در این تحقیق به دنبال ارائه راهکارهای ساده‌تر و عملی هستیم، دو راهکار ساده‌تر را معرفی می‌کنیم:

۷-۵-۱- کمترین میدان در فواصل زمانی مشخص

یکی از راهکارها این است که کمترین میدان[1] در یک مدت زمان مشخص به عنوان سکوت در نظر گرفته شود. به طور مثال در هر ۲۰ تا ۶۰ ثانیه، میدان صدا تحلیل شود و لحظه‌ای که کمترین میدان را دارد به عنوان «سکوت» در نظر گرفته شود.

برای اطمینان می‌توان برای این افت میدان، شرط طول زمانی نیز قائل شد. به عبارت دیگر اگر این میدان برای مدت زمان مشخصی ادامه داشت، سکوت در نظر گرفته شود در غیر این صورت، خیر. مثلاً اگر این افت میدان، حداقل ۲ ثانیه ادامه داشت، یعنی سکوت رخ داده است.

[1] Amplitude

۲-۵-۷- اولین لحظه‌ای که میدان از مقدار مشخصی کمتر شود

راهکار دیگر می‌تواند این باشد که در فواصل زمانی مشخص، به محض اینکه میدان از مقدار مشخصی (فرضاً 20dbA) کمتر شد و برای مدت مشخصی (فرضاً ۲ ثانیه) ادامه داشت، آن لحظه، سکوت در نظر گرفته شود.

۶-۷- یافتن سکوت در کاربرد پخش زنده صدا

در صورتی که نتیجه این تحقیق بخواهد بر روی پخش زنده صدا اعمال شود؛ از جمله در مکالمات تلفنی و یا پخش زنده سخنرانی‌ها، با توجه به اینکه میدان صدا در آینده در دسترس نخواهد بود، راهکار دوم یعنی اولین افت میدان می‌تواند به کار گرفته شود.

اسکریپت‌های مختلفی برای یافتن سکوت در فایل‌های صوتی طراحی شده است که می‌توان به SoX[1] به عنوان یکی از آن‌ها اشاره کرد. این اسکریپت منبع‌باز این امکان را فراهم می‌کند که سکوت در یک فایل صوتی به راحتی تشخیص داده شود.

۷-۷- افزایش به نرمی در کاربرد تبدیل متن به صدا

موضوع دیگر در پیاده‌سازی این ایده، افزایش تدریجی صدا در مدت زمانی است که به سکوت بعد برسیم. این کار می‌تواند با شمارش تعداد ترکیب صامت و مصوت‌ها بین سکوت فعلی و سکوت بعدی و تقسیم کردن عدد ۲۰ (درصد صدایی که در سکوت فعلی کاهش یافته) به آن تعداد شمارش شده، و افزایش حجم صدا در هر ترکیب، پیاده‌سازی شود؛ برای مثال اگر بین سکوت فعلی تا سکوت بعد، ۵۲۰ ترکیب صامت و مصوت قرار داشته باشد، نتیجه‌ی تقسیم ۲۰ بر ۵۲۰، یعنی ۰.۰۳ درصد باید در هر ترکیب صامت و مصوت، حجم صدا افزایش یابد.

۸-۷- افزایش به نرمی در کاربرد صدای آنالوگ ضبط شده

در این کاربرد، با دانستن سکوت بعدی و محاسبه مدت زمان بین سکوت فعلی و سکوت بعدی، می‌توان به صورت دوره‌ای در مدت زمان مشخصی به مرور صدا را افزایش داد. برای مثال اگر بین سکوت فعلی و سکوت بعدی، ۲۲ ثانیه فاصله باشد، می‌توان با تقسیم ۲۰ (یعنی درصد کاهش صدا) به ۲۲ (یعنی طول گام)، عدد ۰.۹ را به دست آورد و در هر ۰.۹ ثانیه یک درصد به حجم صدا افزود.

۹-۷- افزایش به نرمی در کاربرد پخش زنده صدا

در این کاربرد، با توجه به اینکه محل سکوت بعدی مشخص نیست، نمی‌توان از راهکاری که در کاربرد صدای آنالوگ ضبط شده معرفی شده است استفاده نمود. شاید بتوان با «پیش‌بینی سکوت بعدی» راهکاری مانند راهکار قبل ارائه کرد که البته این موضوع، خود می‌تواند یک تحقیق جداگانه باشد، بنابراین آن‌را به فرصتی دیگر و یا به محققینی دیگر واگذار می‌کنیم. اما به عنوان یک راهکار ابتدایی می‌توان از سکوت بعدی صرف نظر کرد و در عوض، از ایده «کاهش در سکوت، افزایش در ۵ ثانیه» استفاده نمود با این تفاوت که زمان ثابت ۵ ثانیه را تا لحظه شروع به کار برای یافتن سکوت بعدی به تأخیر انداخت؛ به طور مثال، پس از تشخیص یک سکوت و کاهش حجم صدا به میزان ۲۰ درصد، می‌توان حجم صدا را طی مدت زمان ۲۰ ثانیه به حالت ۱۰۰ درصد رسانید و سپس شروع به یافتن سکوت بعدی کرد.

۸- شبیه‌سازی و بیان نتایج

برای شبیه‌سازی ایده، ابتدا یک متن نسبتاً طولانی برای تبدیل به گفتار انتخاب شد که فایل صوتی به دست آمده از نرم افزار پارس‌خوان[2] برای این متن، ۳ دقیقه و ۴۵ ثانیه شد.

در انتخاب متن، به این موضوع توجه شد که متن، حاوی یک داستان جذاب باشد تا کاربر را بیشتر مجذوب نماید و از توجه به حواشی اجتناب کند.

برای انجام آزمایشات نیاز داشتیم که ایده‌ی مطرح شده در این پژوهش را بر روی یک فایل صوتی خروجی گرفته شده از یک مبدل متن به گفتار اعمال کنیم و سپس نتیجه را تجزیه و تحلیل نماییم. برای شبیه‌سازی از نرم افزار Corel Video Studio جهت یافتن سکوت‌ها و کاهش حجم صدا و افزایش صدا به نرمی، استفاده شد. این نرم افزار جهت انجام تدوین صدا و تصویر کاربرد دارد.

۱-۸- روش اول اندازه‌گیری: استفاده از دستگاه EEG

در اولین گام با استفاده از دستگاه EEG[3] به ثبت واکنش‌های مغز انسان در هنگام شنیدن یک متن تبدیل شده به گفتار، در دو حالت مختلف؛ یک بار در حالت معمولی و بدون اعمال ایده‌ی مطرح شده در این پژوهش و بار دوم پس از اعمال ایده‌مان پرداختیم.

شکل ۵- یکی از سوژه‌های آزمایش EEG

در این آزمایش یک نوجوان ۸ ساله انتخاب شد تا مطمئن باشیم که حواس او به طور کامل به داستان متوجه است. استفاده از یک

[1] http://sox.sourceforge.net/

[2] Parskhan: http://parskhan.cc

[3] ElectroEncephaloGram

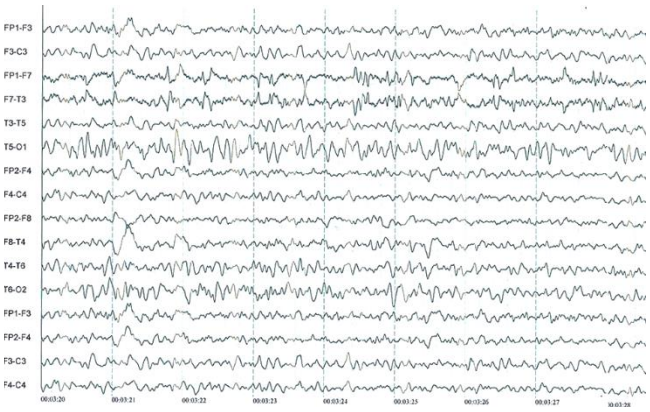
بزرگسال می‌توانست این ریسک را به همراه داشته باشد که درگیری‌های ذهنی وی باعث تغییر گراف EEG گردد.

ابتدا صدای اول (بدون اعمال ایده) توسط هدفون‌هایی که به لپ‌تاپ متصل بود، برای سوژه پخش شد. دلیل پخش صدای معمولی در نوبت اول، این بود که اجازه دهیم ذهن او بر اثر شنیدن صدای طولانی کمی خسته شود تا در کلیپ صوتی دوم، تأثیر ایده را بهتر مشاهده کنیم.
بلافاصله بعد از اتمام کلیپ اول، کلیپ دوم که ایده‌ی این پژوهش بر روی آن اعمال شده بود، پخش شد.

در پایان، ۶۳ صفحه نوار مغز به دست آمد که ۳۱ صفحه مربوط به کلیپ اول، ۱ صفحه مربوط به وقفه‌ای که تا پخش کلیپ دوم ایجاد شد و ۳۱ صفحه مربوط به کلیپ دوم بود.

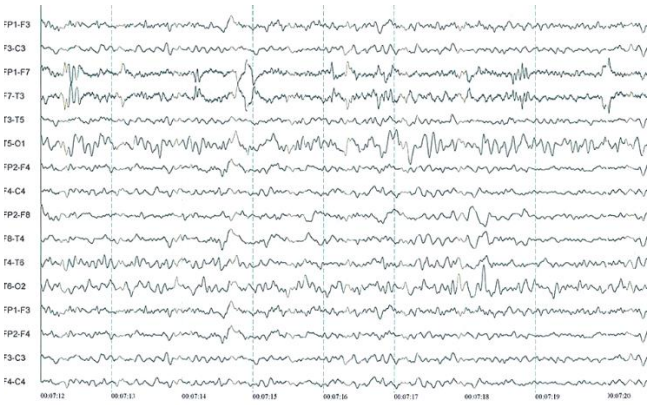
ما ۴۵ ثانیه‌ی آخر از هر دو کلیپ را جدا کرده و در نرم افزار فتوشاپ بر روی یکدیگر قرار دادیم.

یک نمونه‌ی ۸ ثانیه‌ای از کلیپ اول:



شکل ۶ - بخشی از گراف **EEG** مربوط به کلیپ اول

یک نمونه‌ی ۸ ثانیه‌ای از کلیپ دوم:



شکل ۷ - بخشی از گراف **EEG** مربوط به کلیپ دوم

۸-۲- نتیجه اندازه‌گیری به روش اول

نوار مغزهای گرفته شده به سه متخصص مغز و اعصاب نشان داده شد تا شاید مهر تأییدی بر آرامش بیشتر در هنگام شنیدن کلیپ دوم زده شود، اما هر سه بیان داشتند که امکان تشخیص چنین تغییر حسی اندکی در نوار مغزهای EEG وجود ندارد و نیاز به دستگاه‌های بسیار حساس‌تر و دقیق‌تر است که در دسترس همگان نیست.

بنابراین تصمیم گرفتیم به روش دیگری به بررسی ایده‌مان بپردازیم.

۸-۳- روش دوم اندازه‌گیری: نظرسنجی

در این مرحله، یک نظرسنجی تحت وب طراحی کردیم که در آن کلیپ‌ها برای شرکت کنندگان در نظرسنجی پخش می‌شد و سپس از افراد خواسته می‌شد که کلیپی که بهتر تشخیص می‌دهند را انتخاب کنند و دلیل انتخاب خود را توضیح دهند.

برای رفع هر گونه ابهام، در نظرسنجی در مورد الگوریتم خود توضیحی ندادیم. یعنی گفته نشد که کلیپ دوم چه بهبودی نسبت به اولی داشته. حتی توضیح ندادیم که کدام کلیپ از نگاه ما بهتر است و حتی برای اینکه کاربر در ذهن خود تصور نکند که لابد کلیپ دوم بهبود یافته است و آن‌را انتخاب کند، اینطور القا کردیم که جایگاه کلیپ‌ها رندوم است اما در عمل کلیپ معمولی را ابتدا پخش کردیم و کلیپ بهبود یافته را در جایگاه دوم.

از کاربران خواسته شد که در صورت امکان از هدفون برای شنیدن کلیپ‌ها استفاده کنند.

برای اینکه از تقلب و پاسخ‌های غیرمعتبر جلوگیری شود، مدت زمان حضور کاربر در صفحه اندازه‌گیری شد و پاسخ‌هایی که مدت حضور آن‌ها کمتر از مجموع زمان دو کلیپ بود، حذف گردید. همچنین، مشخصات سیستم و آی.پی هر کاربر نیز ثبت شد و مواردی که احتمال داده شد که یک کاربر به صورت تکراری و با نام‌های مختلف در نظرسنجی شرکت کرده حذف گردید.

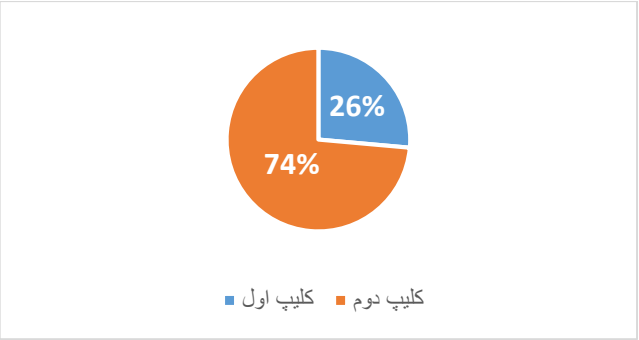
مواردی که کاملاً مشخص بود که معتبر نیست؛ مانند کاربری که به خاطر سرعت کم اینترنت و بافر کردن صدا، دلیل را قطع و وصلی کمتر ذکر کرده بود، حذف گردید.

۸-۴- نتیجه اندازه‌گیری به روش نظرسنجی

از مجموع ۷۷ نظر، ۵۳ نظر معتبر شناخته شد و تحلیل بر روی آن‌ها با استفاده از نرم افزار اکسل[1] شروع شد.

بر این اساس، درصد افرادی که کلیپ اول یعنی کلیپ بدون اعمال ایده را انتخاب کرده‌اند نسبت به کسانی که کلیپ دوم یعنی کلیپ پس از اعمال ایده را انتخاب کرده‌اند، در نمودار زیر مشخص است:

[1] Microsoft Excel



همانطور که مشخص است، به طور شگفت‌آوری، ۷۴ درصد شرکت‌کنندگان، کلیپ دوم را ترجیح داده‌اند.

برخی از دلایل که در نظرات ثبت شده است، کاملاً مشخص می‌کند که ایده این پژوهش به خوبی توانسته باعث بهبود صدای تبدیل شده از متن شود؛ به طور مثال:

- مهدی، دانشجو، ۱۹ ساله گفته است:

«کلیپ دوم مکث دارد و باعث میشه گوش رو اذیت ندهد.حرف زدن گوینده یکنواخت نیست ومیشه تمرکز کرد.و قشنگ میشه حرف گوینده رو فهمید.»

- محمد رضا، کارمند، ۳۲ ساله نظر جالبی دارد که نکات جالبی را مشخص می‌کند:

«با سلام .من هر دو را گوش دادم ولی کلیپ اول خسته کننده بود انگار فقط قراره این متن رو تند تند بخونه و تمام بشه ولی کلیپ دوم با آرامش خاصی و مکث لازم زمانی که جمله تمام میشود وجود دارد که عجله نکردن روی تمام کردن متن این نوع بیان را شیوا تر و متناسب با نوع ادبی متن قرار داده است.»

- وحید، دانشجو، ۲۸ ساله، همان نظری را دارد که ما در مورد بسیاری از مبدل‌های متن به گفتار داریم:

«در کلیپ اول احساس کردم کسی دنبال این بنده خدا میکرد و تندتر میخواند؛ولی ددر کلیپ دوم تقریبا این احساس کمتر بود.»

نگارنده: این نظر زمانی جالب می‌شود که بدانیم ما در کلیپ دوم، سرعت تلفظ را هرگز تغییر ندادیم!

۹- نتیجه‌گیری

بر اساس آن‌چه گفته شد، می‌توان به این نتیجه رسید که ایده‌ی مطرح شده در این پژوهش، یعنی کاهش حجم صدا در سکوت و افزایش حجم صدا به نرمی، می‌تواند تا حد زیادی از خستگی ذهنی کاربر بر اثر شنیدن طولانی مدت صدای یکنواخت به ویژه صدای یکنواخت تولید شده از نرم افزارهای تبدیل متن به گفتار، بکاهد.

البته این موضوع را می‌شود در رفتار روزانه انسان‌ها نیز مشاهده کرد. به طور مثال هنگامی که پس از شنیدن طولانی مدت صدای یکنواخت مجری اخبار، ترجیح می‌دهند کمی حجم صدای تلویزیون را کاهش دهند و سپس ناخواسته پس از مدتی حجم صدا را افزایش می‌دهند و این روال چندین بار ممکن است تکرار شود. ما قصد داشتیم در این پژوهش ثابت کنیم که این رفتار ناخودآگاه می‌تواند در نرم افزارهای پخش صدا و یا نرم افزارهای تبدیل متن به صدا به صورت آگاهانه انجام شود و نیازی نباشد که کاربر خود اقدام به کاهش و افزایش صدا نماید.

ایده‌ی مطرح شده در این پژوهش می‌تواند در کاربردهای بسیاری مورد استفاده قرار گیرد. می‌توان روزی را تصور کرد که یک نرم افزار مبدل متن به گفتار بتواند بدون ایجاد خستگی ذهنی اقدام به خواندن مقالات و داستان‌های طولانی برای مخاطبان کند، از تأثیر منفی صدای یکنواخت مجری اخبار در رادیو و تلویزیون کاسته شود، در نرم افزارهای پخش صدا دکمه‌ای تعبیه گردد که کاربران به محض مواجه شدن با یک سخنرانی یکنواخت با فعال کردن آن دکمه، صدای سخنران را قابل‌تحمل‌تر کنند.

مراجع

[۱] م. آریانا، *مستندات نرم افزار آریانا*، ۱۳۸۹.

[۲] م. م. همایونپور، *پژوهشنامه تبدیل متن به گفتار*، شورای عالی انقلاب فرهنگی، دبیرخانه شورای عالی اطلاع رسانی، ۱۳۹۱.

[۳] "استخراج ابعاد نیازمندی‌های تبدیل متن به گفتار در ایجاد پیکره‌های متناظر زبان فارسی،" شورای عالی اطلاع رسانی، ۱۳۸۸.

[۴] م. م. همایونپور، *ارائه داده ها، دستورالعمل و نرم افزارهای ارزیابی عملکرد سیستم های تبدیل متن به گفتار فارسی*، دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی امیرکبیر، ۱۳۸۶.

[5] J. Treasure, "The 4 ways sound affects us," in *TED*, 2009.

[6] B. Greenfield, "How You Can Use Sound And Music To Change Your Brain Waves With Laser Accuracy And Achieve Huge Focus And Performance Gains," [Online]. Available: http://www.bengreenfieldfitness.com/2012/05/.