سنتز گفتار فارسی با استفاده از فرکانس گام در نرم‌افزار Flite

فاطمه نعیمی[۱]، دانشجوی دکتری؛ وحید قدس[۲]، استادیار

۱- گروه مهندسی برق - واحد سمنان - دانشگاه آزاد اسلامی - سمنان - ایران - f.naimi@stu.semnaniau.ac.ir
۲- باشگاه پژوهشگران جوان و نخبگان - واحد سمنان - دانشگاه آزاد اسلامی - سمنان - ایران - v.ghods@semnaniau.ac.ir

چکیده: در این پژوهش، طراحی و پیاده‌سازی سنتز کننده گفتار به زبان فارسی با استفاده از نرم‌افزار Flite ارائه شده است. بدین طریق که ابتدا توسط پایگاه داده جملات فارسی فارس‌دات[۱]، میانگین و انحراف معیار[۲] فرکانس گام و بزرگترین فرکانس گام هر واج[۳] واک‌دار را به‌دست می‌آوریم. سپس تغییرات هر واج عبارت موردنظر را از طریق محاسبه مقدار ارزش[۴] آن با توجه به فرمول ارائه شده پیشنهادی، در نرم‌افزار وارد می‌کنیم. ویژگی این سنتز کننده، تبدیل متن به گفتار با لهجه و تلفظ فارسی می‌باشد. در انتهای این مقاله، نتایج حاصل از اجرای الگوریتم پیشنهادی با منحنی فرکانس‌های گام به‌دست آمده توسط پایگاه داده جملات فارسی فارس‌دات، مقایسه شده است. همچنین، در این پژوهش چند نمونه از جملات زبان فارسی مربوط به پایگاه داده جملات فارسی فارس‌دات، با استفاده از روش پیشنهادی بر روی نرم‌افزار Flite، مورد سنتز قرارگرفته است. آزمون‌های شنیداری، برای میزان قابل‌فهم بودن، طبیعی بودن و خوشایند بودن جملات مذکور انجام شده است که نتایج آن‌ها برای جملات آموزش به‌ترتیب ۴/۴، ۴/۲ و ۴/۶ می‌باشد. همچنین برای جملات مجموعه آزمون، به‌ترتیب برابر ۴/۲، ۴/۱ و ۴/۳ می‌باشد.

واژگان کلیدی: تبدیل متن به گفتار، سنتز گفتار، فارسی، فرکانس گام، نرم‌افزار Flite.

Farsi Speech Synthesis Using Pitch Frequency in Flite software

F. Naiemi[1], PhD Student; V. Ghods[2], Assistant Professor

1-Electronic Group, Semnan Branch, Islamic Azad University, Semnan, Iran, Email: f.naimi@stu.semnaniau.ac.ir
2-Young Researchers and Elite Club, Semnan Branch, Islamic Azad University, Semnan, Iran, Email: v.ghods@semnaniau.ac.ir

Abstract: This survey introduces a model and the implementation of a speech synthesizer in Farsi language using Flite software. In this approach, the mean and the standard deviation of pitch frequency of each voiced phoneme are first calculated by a database of Farsi sentences (Fars Dat). Then, the changes of each phoneme of the desired phrase are inserted into the software through the calculation of a value. The main feature of this synthesizer is its ability to change text to speech within Farsi pronunciation and in Farsi dialect. At the end of this paper, the results of this algorithm are compared to the changes of pitch frequencies extracted from the database of Farsi sentences. Some examples of the sentences from the database are also synthesized using our proposed method on Flite Software. The value of MOS test for understandability, naturalness and good sounding of those sentences are 4.4, 4.2, and 4.6 for the training set, respectively, and 4.2, 4.1, and 4.3 for the test set, respectively.

Keywords: Text to speech, speech synthesizer, Farsi (Persian), pitch frequency, Flite software.

تاریخ ارسال مقاله: ۱۳۹۶/۰۵/۱۵
تاریخ اصلاح مقاله: ۱۳۹۶/۰۸/۲۰
تاریخ پذیرش مقاله: ۱۳۹۶/۰۸/۲۶
نام نویسنده مسئول: وحید قدس
نشانی نویسنده مسئول: ایران - سمنان - دانشگاه آزاد اسلامی - واحد سمنان - باشگاه پژوهشگران جوان و نخبگان.

۱- مقدمه

تصور اینکه یک ماشین بتواند گفتار تولید نماید سال‌هاست که با بشر بوده است، اما تحقق وجود چنین ماشین‌هایی در طی ۵۰ سال گذشته محقق شده است. یکی از اولین کاربردهای سنتز گفتار در سال ۱۹۳۶ اتفاق افتاد. هومر دادلی [۱] در آزمایشگاه‌های بل، یک ابزار مکانیکی تولید کرد که از طریق حرکت پدال‌ها و کلیدهای مکانیکی کار می‌کرد. این ماشین می‌توانست در صورت وجود یک کاربر آموزش‌دیده، صداهایی تولید نماید (اگر داده‌های مناسبی به آن می‌رسید) که مانند گفتار به نظر آیند. این دستگاه که Voder نام داشت برای اولین بار در نمایشگاهی بین‌المللی در سال ۱۹۳۹ در نیویورک و سانفرانسیسکو به نمایش گذاشته شد [۱].

تحقق این موضوع که سیگنال‌های صوتی می‌توانند به‌عنوان مدل های فیلتر- منبع[۵] تجزیه شوند، به‌طوری‌که فاصله بین تارهای صوتی مانند یک منبع صدا و قسمت‌های مختلف دهان مانند فیلترهایی در نظر گرفته شوند، در ابزارهای الکترونیکی آنالوگ مورداستفاده قرار گرفت که می‌توانستند گفتار انسانی را تقلید نمایند. Voder توسعه‌یافته که مجدداً توسط هومر دادلی ساخته شد، می‌توانست نمونه‌ای از این ابزارها باشد [۱].

رشد سنتز الحاقی[۶] در دهه ۷۰ آغاز شد. در سال ۱۹۷۲ راهنمای استاندارد Unix دستورهایی برای پردازش متن به گفتار، تشکیل تجزیه و تحلیل متن، پیش‌بینی آوایی، تولید واج و سنتز شکل موج از طریق یک سخت‌افزار ویژه را ارائه نمود [۲]. ابزارهای speak ‘n spell تگزاس که در اواخر دهه ۷۰ منتشر شدند، یکی از نمونه‌های اولیه تولید انبوه سنتزکننده‌های گفتار بودند [۱]. کیفیت این ابزارها در مقایسه با استانداردهای کنونی بسیار ضعیف بود؛ اما در آن زمان بسیار جالب‌توجه بودند. گفتار، اساساً با استفاده از کد کردن مبتنی بر پیش‌بینی خطی[۷] کدگذاری می‌شد و اغلب از کلمات و حروف جدا از هم استفاده می‌کرد. سنتزگر MITalk متعلق به دنیس کلات از بسیاری جهات مفهوم سنتز گفتار خودکار را به جهان معرفی نمود [۱]. این ابزار که بعداً به محصول DECTalk [۱] ارتقا یافت، گفتاری رباتی اما بسیار قابل‌فهم تولید می‌نمود. این ابزار یک سنتزگر فرمنت بود که تمام صنعت و هنر آن زمان را در خود جای داده بود.

در اواخر دهه ۸۰ سنتزگرهایی که فقط نرم‌افزاری بودند نیز تولید شدند، کیفیت گفتار هنوز هم شبیه صدای انسان نبود، اما گفتار می‌توانست تقریباً لحظه‌ای تولید شود. با ظهور ماشین‌های سریع‌تر و با فضای ذخیره‌سازی بیشتر، مردم شروع به ارتقای سنتزکننده‌ها با استفاده از ذخیره‌های بزرگ‌تر و متنوع‌تر برای گفتار الحاقی نمودند. یوشینوری سگیساکا [۲] در مرکز تحقیقات پیشرفته ارتباطات از راه دور ژاپن ابزار Nuu-talk را در اواخر دهه ۸۰ و اوایل دهه ۹۰ تولید نمود. این ابزار ذخیره بسیار بزرگ‌تری از واحدهای الحاقی ارائه می‌نمود. در نتیجه به‌جای یک نمونه برای هر واحد دو واجی، نمونه‌های بسیاری در اختیار بود. حتی تا سال ۱۹۹۴ زمان موردنیاز برای تولید فایل‌های پارامتری برای یک صدای تازه در Nuu-talk (۵۰۳ جمله) تقریباً چند روز کاری یک پردازنده بود و سنتز به‌صورت لحظه‌ای امکان‌پذیر نبود.

کورزویل [۳] پیش‌بینی کرد نسبت قیمت به عملکرد در سال ۲۰۰۵ باعث تولید سیستم‌های سنتز گفتار ارزان‌تر و قابل‌دسترس بیشتر شده و بیشتر مردم می‌توانند از این سیستم بهره‌مند شوند، همچنین از سال ۲۰۰۵ به بعد برخی از محققان با استفاده از مجموعه داده مشترک گفتار به ارزشیابی سیستم‌های سنتز گفتار پرداختند [۳].

در دسترس بودن سیستم‌های سنتز رایگان و نیمه‌رایگان مانند سیستم سنتز گفتار Flite [۴] و پروژه MBROLA [۵] هزینه‌های ورود به زمینه سنتز گفتار را کاهش می‌دهد و گروه‌های بیشتری در ارتقای این سیستم‌ها مشارکت می‌نمایند.

ازجمله تلاش‌های محققان ایرانی در مورد موضوع تبدیل متن به گفتار فارسی می‌توان موارد زیر را ذکر نمود:

۱-تبدیل حرف به صدا در زبان فارسی به کمک شبکه‌های عصبی پرسپترون چندلایه‌ای [۶] که به دلیل عدم استفاده از اعراب در نوشتار و درنتیجه مستور بودن بعضی از واکه‌ها، یک سیستم تبدیل حرف به صدا با معماری سه لایه بررسی شده است. لایه اول این سیستم قانون-گرا می‌باشد و لایه دوم از پنج شبکه عصبی پرسپترون چندلایه‌ای و یک بخش کنترل‌کننده برای تعیین دنباله واجی متناظر با حروف تشکیل شده است.

۲-تولید پارامترهای سنتز گفتار فارسی با استفاده از مدل‌های مخفی مارکوف و درخت تصمیم‌گیری [۷]. در مقاله [۷]، برای پیاده‌سازی سیستم سنتز، از مدل‌های مخفی مارکوف برای مدل‌کردن پارامترهای مربوط به واحدهای گفتاری استفاده شده است و برای تبدیل ضرایب کپسترال به سیگنال صحبت، از فیلتر MLSA استفاده شده است. برای استخراج فرکانس گام، روش خودهمبستگی اصلاح‌شده مورد استفاده قرار گرفته است. برای تولید پارامترهای سنتز گفتار توسط HMMها، از الگوریتمی استفاده شده که در آن، برای در نظر گرفتن اطلاعات بافت، علاوه بر ویژگی‌های ضرایب کپستروم و فرکانس گام، مشتق اول و دوم آن‌ها نیز، مورد استفاده قرار گرفته است [۷]. برای به‌دست آوردن مدل طول زمانی واج‌ها، مشاهدات موجود از هر ترایفون را در پایگاه داده، طبق الگوریتم ویتربی با مدل HMM آن مقایسه نموده و دنباله حالات طی شده را به‌دست آورده و با میانگین‌گیری از تعداد دفعات حضور در هر حالت مدل HMM ترایفون، متوسط طول زمانی حضور در هر حالت را برای هر ترایفون به‌دست آمده است [۷]. زمان‌های میانگین حاصل، مدل‌های طول زمانی برای هر ترایفون را تشکیل می‌دهند. در هنگام سنتز با توجه به مدل طول هر حالت از مدل HMM هر ترایفون، پارامترهای هر کدام از حالت‌های HMM، شامل بردار میانگین و بردار واریانس آن حالت تکرار و با استفاده از این پارامترها، دنباله ضرایب کپسترال و گام موردنیاز برای سنتز گفتار به‌دست آمده و توسط فیلتر MLSA به گفتار تبدیل

شده‌اند [۷]. برای در نظر گرفتن تأثیر پارامترهای مختلف بر نحوه تلفظ آواها، از درخت‌های تصمیم‌گیری CART، استفاده شده است. این درخت‌ها نقش تولید گام و طول زمانی واج‌ها را بر عهده دارند. برای تولید خودکار فرکانس گام از روش مطرح شده توسط فوجی ساکی [۷] استفاده شده است. در این روش برای مدل کردن شکل کلی سیگنال گام، یک جزء عمومی و برای مدل کردن تأکیدها، تعدادی اجزای محلی در نظر گرفته‌شده است [۷].

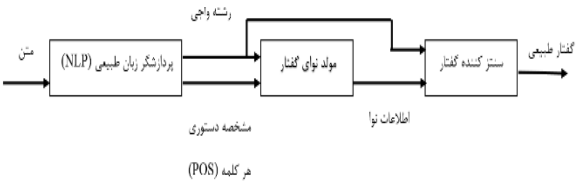
۳- تبدیل حرف به صدا در سیستم‌های تبدیل متن به گفتار فارسی با استفاده از درخت‌های تصمیم‌گیری CART [۲] که در آن پارامترهای استخراج‌شده از کلمات متن آموزش داده شده و سپس از درخت در تعیین آواهای تشکیل‌دهنده کلمات استفاده شده است [۲].

در ادامه مقاله، در بخش ۲ توضیحاتی در خصوص سنتزکننده گفتار ارائه شده است. در بخش ۳، تبدیل متن به گفتار در نرم‌افزار Flite بررسی شده است. بخش ۴ مدل پیشنهادی برای تولید گفتار به زبان فارسی با استفاده از نرم‌افزار Flite را معرفی می‌کند. در بخش ۵، نتایج حاصل از مدل پیشنهادی بر روی جملات فارسی ارائه شده و در پایان، بخش ۶ به نتیجه‌گیری از مباحث پرداخته است.

۲- سیستم سنتز گفتار

سنتز گفتار یک فناوری است که به‌وسیله آن متن به گفتار مصنوعی[۸] تبدیل می‌شود [۸]. در موضوع سنتز گفتار، ذخیره‌سازی تمام کلمات یک زبان غیرممکن (و اغلب بی‌فایده) است. در واقع سنتز گفتار، تولید گفتار از طریق رونویسی حروف به آوا، به‌منظور گفتن کلمات و جملات می‌باشد.

شکل ۱ روند بنیادی یک سنتز کننده گفتار را نشان می‌دهد [۹].



شکل ۱: سیستم تولید متن به گفتار فارسی [۹].

۲-۱ پردازش‌گر زبان طبیعی[۹]

در بلوک پردازش‌گر زبان طبیعی جملات ورودی به فهرستی از کلمات تبدیل می‌گردند [۱۰]. در این راستا اعداد، حروف اختصاری، علائم و نشانه‌ها نیز به معادل کامل متنی خود تبدیل می‌شوند. سپس بر روی کلمات جمله، تحلیل شکلی[۱۰] انجام می‌شود (در زبان فارسی و انگلیسی تطابق مستقیمی بین حروف و واج وجود ندارد، به‌عنوان مثال در زبان فارسی می‌توان به موارد زیر اشاره نمود: وجود یک حرف برای چند واج، استفاده از حروفی که خوانده نمی‌شوند، نوشته نشدن واکه‌های کوتاه، نوشته نشدن کسره اضافه) و مشخصه دستوری هر کلمه[۱۱] تعیین می‌گردد. در مرحله بعد با توجه به مشخصات دستوری کلمات، هر کلمه به رشته واجی مربوط به خود تبدیل می‌شود. یکی از بدیهی‌ترین راه‌ها برای تولید رشته واجی، مجموعه قواعد تبدیل حرف به صدا می‌باشد که حروف متن را به رشته‌ای از واج‌ها تبدیل می‌نماید.

۲-۲ مولد نوای گفتار

۲-۲-۱ نوا[۱۲]

نوا یکی از فاکتورهای اصلی برای به‌دست آوردن یک گفتار مصنوعی باکیفیت زیاد می‌باشد. مفهوم نوا، زیروبم کردن صدا و لحن گفتار است که باعث تلفظ و برداشت مفهوم‌های مختلفی از گفتار می‌شود. همچنین اطلاعاتی راجع به وضعیت روانی سخن‌گو ارائه می‌دهد. در صدای هر شخص بخش‌های صدادار[۱۳] و بخش‌های بی‌صدا[۱۴] وجود دارد. بخش‌های صدادار صدای هر شخص، دارای یک فرکانس اصلی می‌باشد که بر اساس آن تارهای صوتی به لرزه درمی‌آیند. می‌توان گفت فرکانس اصلی یکی از فاکتورهای جداشونده صدا می‌باشد که قابل‌تغییر است. به‌طور ساده، فرکانس اصلی به‌وسیله اصلاح شکل‌موج قابل‌تغییر می‌باشد.

مفهوم مولد نوای گفتار زیروبم کردن صدا یا تغییرات فرکانس گام [۱۳-۱۱]، دیرش[۱۵] [۱۶-۱۴]، شدت[۱۶] [۱۷، ۱۸] و درنگ[۱۷] [۱۹] می‌باشد.

۲-۳ سنتز گفتار

در خصوص بلوک سنتز گفتار باید بدانیم سنتزگر گفتار مبتنی بر سه روش می‌باشد:

۱- سنتز شمرده به شمرده لغات[۱۸]

۲-سنتز فرمنت[۱۹]

۳-سنتز الحاقی[۲۰]

۲-۳-۱ سنتز شمرده به شمرده لغات

در این روش از یک مدل کامپیوتری‌ـمکانیکی برای تولید گفتار استفاده می‌شود [۹، ۲۰]. مثل حنجره که به‌وسیله ارتعاش تارهای صوتی باعث تولید صدا و تحریکات حلقی می‌شود. زبان به عنوان بخشی از فیلتر مسیر صوتی[۲۱] بوده و در ایجاد اصوات نقش دارد. مسیر عبور هوا به‌وسیله زبان با فشار بر روی سقف دهان یا دندان‌ها تغییر می‌کند. لب ها برای ایجاد هماهنگی و کمک به فرایند تولید گفتار به زبان کمک می‌کنند. در حالت ایده‌آل سنتزکننده‌های شمرده به شمرده می‌توانند به‌وسیله ماهیچه‌های مصنوعی و کنترل آن‌ها به شکل زبان، لب و دهان شبیه‌سازی شوند.

۲-۳-۲ سنتز فرمنت

در دهه‌های اخیر بیشترین استفاده در متن روش‌های سنتز، روش سنتز فرمنت بوده است که بر اساس مدل فیلتر-منبع از گفتار

پایه‌ریزی شده است [۹، ۲۰]. حداقل سه فرمنت برای تولید یک گفتار قابل مفهوم لازم است و اگر به پنج عدد برسد می‌توان گفتار باکیفیت بالا ایجاد کرد. هر فرمنت معمولاً با یک تشدیدکننده دوقطبی که قادر است هم فرکانس فرمنت و هم پهنای‌باندش را مشخص کند، مدل می‌شود.

۲-۳-۳ سنتز الحاقی

در دو روش قبل پارامترهای مشخصه گفتار در هر بازه زمانی توسط مجموعه‌ای از قواعد تولید می‌شدند، اما در این روش واحدهای گفتار ذخیره‌شده طبیعی برای تولید گفتار خروجی به‌صورت تکه‌تکه در کنار هم قرار می‌گیرند [۹، ۲۰]. یکی از مهم‌ترین جنبه‌ها در سنتز الحاقی انتخاب طول واحد صحیح است. دو مشکل در سنتز الحاقی نسبت به دیگر روش‌ها موجود است:

- بروز ناپیوستگی در نقاط اتصال که با استفاده از دوواجی‌ها و روش‌های خاصی برای هموارسازی سیگنال می‌توان آن را حل کرد.
- نیاز به حافظه بالا مخصوصاً زمانی که طول واحدها هم‌چون کلمه‌ها یا هجاها باشد.

۳- نحوه کار و تبدیل متن به گفتار در نرم‌افزار Flite

۱-۳ مقدمه

Flite که هسته اصلی آن از سیستم سنتزکننده گفتار Festival [۵] گرفته شده است، یک ماشین کوچک سنتزکننده گفتار زبان انگلیسی با زمان اجرای سریع است که برای سیستم‌های تعبیه‌یافته و سرورها مناسب می‌باشد [۴]. این ماشین که ابزارها، الگوها و مدارک موردنیاز برای ساخت یک صدای ساختگی (مصنوعی) جدید را فراهم می‌آورد، در ANCI C نوشته شده است و طوری طراحی شده که تقریباً به هر پایگاهی قابل‌انتقال بوده و دارای سخت‌افزار بسیارکوچکی است. درواقع Flite یک کتابخانه سنتز است که می‌تواند در داخل دیگر برنامه‌ها قرار گیرد.

۲-۳ سنتز قطعات گفتار در Flite

واحد گفتار در این سنتز کننده، الحاقی از نوع PSOLA TD[۲۲] است که درواقع الگوریتم اضافه کردن هم‌پوشانی گام‌های هم‌آهنگ در حوزه زمان می‌باشد.

در این روش از علائم گام[۲۳] به‌جای دوره تناوب گام استفاده می‌شود. علائم گام، گام‌هایی هستند که بزرگ‌ترین قله را در هر دوره تناوب دارند. برای تعیین علائم گام ابتدا شکل موج را توسط دو فیلتر بالاگذر و پایین‌گذر (یک فیلتر میان‌گذر) با توجه به جنسیت گوینده محدود می‌کنیم. در این نرم‌افزار به‌طور قراردادی حد پایین ۰/۰۰۵ ثانیه و حد بالا ۰/۰۱۲ ثانیه متعلق به یک گوینده مرد (محدوده فرکانسی ۸۰ تا ۲۰۰ هرتز) انتخاب شده است. با توجه به اینکه صداهای بی‌واک دوره تناوب ندارند، علائم گام را به‌صورت فرضی برای صداهای بی‌واک (غیر از سکوت) در هر ۱۰ میلی‌ثانیه تعیین می‌کنیم. بدین‌صورت که یک منحنی گام به‌وسیله درون‌یابی صداهای واک‌دار مجاور، به صدای بی‌واک (غیر از سکوت) اختصاص داده و نتیجه آن هموارسازی می‌شود؛ بنابراین در این روش سکوت از گفتار جدا می‌شود اما راهی برای تعیین علائم گام صداهای واک‌دار از علائم گام صداهای بی‌واک وجود ندارد [۴].

تعیین تغییرات فرکانس گام در این نرم‌افزار از طریق تعریف میانگین و انحراف معیار می‌باشد؛ که از طریق رابطه ۱ به‌دست می‌آید:

$$map\_f0(v,m,s)=((((v-mm)/ms)*s)+m) \quad (۱)$$

که در آن v=value ،m=mean ،s= stddev ،mm=model_mean =۱۷۰ و ms=model_stddev= ۳۴ است.

در این برنامه، مقادیر model_mean و model_stddev ثابت می‌باشند.

در این برنامه تنها می‌توانیم با تغییر میانگین و انحراف معیار که برای تمامی واج‌ها یکسان تعریف شده است، منحنی فرکانس گام تمام واج‌ها را به مقدار یک‌سان تغییر دهیم؛ اما نمی‌توانیم زیروبمی هر واج را جداگانه تغییر دهیم. بنابراین صدای به‌دست‌آمده بسیار رباتیک می‌باشد و فقط با تغییر عدد مربوط به میانگین و انحراف معیار که برای کلیه واج‌ها می‌باشد، تغییرات یک‌سان به‌کار می‌رود.

واحد پایگاه داده در این سنتز کننده، دوواجی‌ها با حجم ۲/۱ مگابایت می‌باشند. در این پایگاه داده آدرس مربوط به هر یک از دوواجی‌ها مقابل آن نوشته شده است و برای هر یک پارامترهای مربوط به ضرایب LPC و سیگنال باقی‌مانده[۲۴] به‌دست می‌آید. با توجه به نرخ نمونه‌برداری، مرتبه ضرایب LPC در این نرم‌افزار ۱۰ می‌باشد، همچنین فرکانس نمونه‌برداری ۸ کیلوهرتز و نمونه‌ها ۱۶ بیتی هستند. Flite شامل ۴ بخش اصلی می‌باشد [۴]:

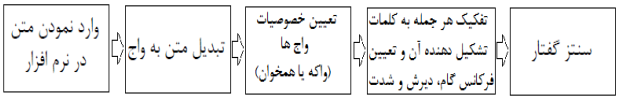
۱. کتابخانه Flite: شامل کدهای مرکزی است. این قسمت از سیستم CST نامیده می‌شود که برای ابزارهای گفتار در C به‌کار می‌رود.
۲. مدل زبان: مجموعه‌ای از آواها، قوانین گفتاری، آنالیز متن و مدل‌های نوایی را فراهم می‌آورد که اغلب در بازشناسی گفتار مورد استفاده قرار می‌گیرد. این قسمت از سیستم US نامیده می‌شود.
۳. مدل لغت‌نامه: یک مدل طرز تلفظ و ادای سخن بوده که شامل یک لغت‌نامه و قانون تبدیل حروف به صدا برای بیرون کشیدن الفاظ کلمات می‌باشد. لغت‌نامه‌ها به واحدهای ذخیره‌شده زبان وابسته هستند، این قسمت از سیستم CMU نامیده می‌شود. از نظر اندازه در نرم‌افزار Flite می‌توان گفت که لغت‌نامه‌ها و سیستم‌های قانون تبدیل حروف به صدا بیشترین حافظه نرم‌افزار را (در زبان انگلیسی) تشکیل می‌دهند.
۴. صدا: شامل فهرستی از مدل‌های نوایی یک گوینده خاص و تعریف صدای او می‌باشد. موارد ابتدایی در هر صدا توسط مدل زبان مربوط به آن، فراهم می‌شود. در این نرم‌افزار فایل صوتی با فرمت

wav، با مشخصات مونو،۱۶ بیت، ۸ کیلوهرتز pcm بوده و گوینده مرد می‌باشد.

۴- مدل پیشنهادی برای تولید گفتار به زبان فارسی

این تحقیق با استفاده از برنامه سنتزکننده گفتار جامع و کامل Flite انجام شده است. ابتدا به منبع برنامه دسترسی پیدا کرده و زیربرنامه‌های موردنیاز در قالب C در کنار هم متصل شده است. سپس با تغییرات انجام‌شده و اضافه نمودن اطلاعات موردنیاز بر روی این زیربرنامه‌ها، سنتزکننده گفتار در زبان فارسی تهیه شده است.

در شکل ۲ روندنمای تبدیل متن به گفتار نرم‌افزار Flite نشان داده شده است که در قسمت مربوط به واج‌ها، اطلاعات فرکانس گام مربوط به زبان فارسی وارد شده است.



شکل ۲: روندنمای تبدیل متن به گفتار در نرم‌افزار Flite.

بدین‌طریق که در بلوک وارد نمودن متن در نرم‌افزار، ابتدا در قسمت اصلی و فراخوان زیربرنامه، نام فایلی را که نرم‌افزار باید متن را از آن بخواند وارد می‌کنیم.

همچنین نام فایلی (با فرمت Wave) را که می‌خواهیم متن خوانده‌شده در آن قرار گیرد، به برنامه اضافه شده است.

در بلوک بعدی استخراج صورت واجی متن انجام می‌شود. برای مثال عبارت «دیروز برف بارید» به واج‌های زیر تبدیل می‌شود (pau به معنای وقفه می‌باشد.):

"diruz barf barid"
pau d ih r uw z b aa r f b aa r ih d pau

به‌طور کلی تبدیل حرف به صدا به دو دسته مبتنی بر فرهنگ لغت و مبتنی بر قانون طبقه‌بندی می‌شود. فرهنگ لغت این نرم‌افزار شامل ۱۰۰۰۰ کلمه می‌باشد. همچنین از قواعد تبدیل حروف به صدا نیز به‌صورت یک پشتیبان برای کلماتی که صریحاً در لغت‌نامه نیامده‌اند، استفاده می‌شود. نرم‌افزار علاوه بر یک فرهنگ‌نامه لغت، از فهرستی به نام افزایش نیز پشتیبانی می‌کند که برنامه‌های مشخص و کاربران آن‌ها را قادر می‌سازد تا ورودی‌های جدیدی را به لغت‌نامه اضافه کنند.

در جدول ۱ حروف فارسی و نشانه و واج‌های معادل انگلیسی نشان داده شده است. همچنین نشانه و واکه انگلیسی معادل ۶ واکه فارسی در جدول ۲ نشان داده شده است. در جدول ۳ تبدیل متن به واج ۴۰ جمله گرفته‌شده از پایگاه داده فارسی نشان داده شده است.

بلوک بعدی مربوط به تعیین خصوصیات واج‌ها می‌باشد. در ادامه برنامه‌های مربوط به این بلوک، خصوصیات هر یک از واج‌ها در قالب نمودار درختی آورده شده است که به‌صورت زیر می‌باشد:

- هر واج به یکی از گروه‌های واکه[۲۵] (+) یا همخوان[۲۶] (-) تقسیم می‌شود. (vc[۲۷] + -)

جدول ۱: حروف فارسی و نشانه‌ها و واج‌های معادل انگلیسی.

واج	حرف (نشانه) انگلیسی	حرف (نشانه) فارسی		واج	حرف (نشانه) انگلیسی	حرف (نشانه) فارسی
/s/	/s/	ص		/aa/	/aa/	آ
/z/	/z/	ض		/b/	/b/	ب
/t/	/t/	ط		/p/	/p/	پ
/z/	/z/	ظ		/t/	/t/	ت
-	-	ع		/s/	/s/	ث
/gh/	/gh/	غ		/j/	/j/	ج
/f/	/f/	ف		/ch/	/ch/	چ
/gh/	/gh/	ق		/h/	/h/	ح
/k/	/k/	ک		/kh/	/kh/	خ
/g/	/g/	گ		/d/	/d/	د
/l/	/l/	ل		/z/	/z/	ذ
/m/	/m/	م		/r/	/r/	ر
/n/	/n/	ن		/z/	/z/	ز
/v/	/v/	و		/zh/	/zh/	ژ
/hh/	/hh/	ه		/s/	/s/	س
/y/	/y/	ی		/sh/	/sh/	ش

جدول ۲: نشانه‌ها و واج‌های انگلیسی معادل ۶ واکه فارسی.

واج	نشانه انگلیسی	مصوت فارسی
/ae/, /aa/, /ax/	/a/	فتحه
/eh/	/e/	کسره
/ao/, /aa/, /ax/, /ow	/o/	ضمه
/ae/, /aa/	/a:/	آ – ا
/ih/, /iy/, /ax/	/ee:/, /i/	ی
/uw/, /uh/, /ow/, /ax/	/u:/	و

- در صورت واکه بودن:

به کدام‌یک از ۴ گروه زیر تعلق دارند؟

کوتاه (s)، بلند (l)، واکه دوگانه[۲۸] (d)، حرف صدادار میان کلمه بدون تشدید[۲۹] (a).

(vlng[30] s l d a ۰)

- در هنگام ادا کشیدگی صدا در چه حدی است. زیاد (۱)، متوسط (۲)، کم (۳)

(vheight ۱ ۲ ۳ -)

- در هنگام ادا کدام‌یک از قسمت‌های جلو (۱)، وسط (۲)، پشت (۳) در تارهای صوتی مشغول می‌شوند.

(vfront[31] ۱ ۲ ۳)

- چگونه ادا می‌شوند. زبان با لب‌ها لمس می‌شود (+)، زبان گرد می‌شود (-)

(vrnd[32] + -)

در صورت هم‌خوان بودن:

- به کدام‌یک از گروه‌های زیر تعلق دارند؟

وقفه‌ها (s)، سایشی‌ها[۳۳] (f)، ترکیب وقفه‌ها و سایشی‌ها[۳۴] (a)، خیشومی‌ها[۳۵] (n)، شبه‌واکه مایع[۳۶] (l)

(ctype[37] s f a n l ۰)

- در هنگام تلفظ کدام قسمت واج‌گاه مشغول می‌شود.

لب‌ها[۳۸] (l)، لب‌ها به‌صورت حباب‌دار جمع می‌شوند[۳۹] (a)، سخت‌کامی[۴۰] (p)، لب‌ها و دندان‌ها (b)، دندان‌ها (d)، نرم‌کامی[۴۱] (v).

(cplace[42] l a p b d v ۰)

- واک‌دار یا بی‌واک هستند.

(cvox[43] + -)

جدول ۳: تبدیل متن به واج.

جمله فارسی	جمله معادل فارسی	تبدیل متن به واج
۱- دیروز برف بارید.	Dinuz barf barid.	Pau d ih r uw z b aa r f b aa r ih d pau
۲- من یک دستگاه چاپ ساخته‌ام.	Man yek datgaaheh chaap saakhteham.	Pau m ae n y eh k d ae st g aa hh ax ch aa p s aa k t ax m pau
۳- پارچ توی کمد است.	Parch tooyeh komod ast.	Pau p aa r ch t uw ax k ow m aa d ae s t pau
۴- کتفم درد می‌کند.	Ket faam dard mikonad.	Pau k eh t f ae m d aa r d m ih k ax n ae d pau
۵- به چه علتی سگ فرار کرد؟	Beh che ellati sag faraar kard?	Pau b eh c hey l aa t iy s ae g faa r k aa d pau
۶- دیروز در کرج، قاسم را دیدم.	Deeruz dar karaj gaasem ra didam.	Pau d ih uw z d aa r k ae r ay g az s ax m r aa d ih d ax m pau
۷- گرگ به گله زد.	Gorge beh galleh zad.	Pau g ao r g b eh g ae l ax z ae d pau
۸- پایش از مچ شکست.	Payash az mohche shekast.	Pau p aa y ax sh ae z m ow ch sh eh k ax s t pau
۹- پاپ فردا می‌آید ایران.	Paap fardaa miaayad eeraan.	Pau p aa p f aa r d aa m iy aa ay ax d eh r aa n pau
۱۰- سوت گم شد.	Soot gome shod.	Pau s uh t g ow m sh ow d pau
۱۱- گاو توی مزرعه است.	Gaav tooyeh mazraeh ast.	Pau g aa v t uw ax m ael z r ey aa s t pau
۱۲- پیچ باز شد.	Pich baaz shod.	Pau p ih ch b aa z sh ow d pau
۱۳- جیپ راه افتاد.	Jip rah ohftaad.	Pau jh ih p r aa hh aa l f t aa d pau
۱۴- آن اسب چه می خورد؟	An asb cheh mikhorad?	Pau ae n ae s b ch ehl m ih k hh ao r ax d pau
۱۵- پمپ، آب ندارد.	Pomp aab nadaarad.	Pau p aa m p ae b n aa d aa aa d pau
۱۶- دو رکعت، نماز خواند.	Doe rakat namaaz khaand.	Pua d ow r ae k ae t n aa m aa z k ax n d pau
۱۷- پول را قاب زدم.	Pule ra gaap zadam.	Pau p uw l r aa g aa p z aa d ax m
۱۸- سرتیپ وقت ندارد.	Sartipe vagt nadaarad.	Pau s aa I t ax p v ae g t n aa d aa aal d pau
۱۹- دیروز توت چیدم.	Deeruz toot chidam.	Pau d ih r uw z t uw t chi h d ax m pau
۲۰- کتابت پاره شده است.	Kehtaabat paareh shodeh ast.	Pau k eh t ax b aa t p er ax sh owl d eh ae s t pau
۲۱- من لوسش نکرده‌ام.	Man lusash nakradeham.	pau m ae n l uw s ax sh n ax k aa r d ax hh ax m pau
۲۲- حجاج به مکه رفتند.	Hojjaaj beh makkeh raftand.	pau hh oy jh aa jh b eh m ae k ax r ae f t ax n d pau
۲۳- پاییز فصل برگ ریزان است.	Paaez fasleh bargrizaan ast.	pau p aa eh z f ae s l ax b aa r g r ih z aa n ae s t pau
۲۴- داس، شکست.	Daas shekast.	pau d aa z sh eh k ax s t pau
۲۵- پارچ لب طاقچه است.	Parch labeh tagcheh ast.	pau p aa r ch l ae b ax t ae g ch ax ae s t pau
۲۶- سگ علی را گاز گرفت.	Sag ali ra gaaz gehreft.	pau s ae g aa l iy r aa g aa z g aa r eh f t pau
۲۷- این ظرف، در ندارد؟	In zarf dar nadaarad?	pau ih n z aa r f d aa r n aa d aa aa d pau
۲۸- فقر باعث یاس می‌شود.	Faghre baaiseh yase mishavad.	pau f ae g er b ae ey s ax y aa s ax m ih sh ax v ax d pau
۲۹- جنب خانه ما یک بانک است.	Janbeh khaneyeh ma yak bank ast.	pau jh ae n b ax k aa n iy ax m aa y eh k b aa ng k ae s t pau
۳۰- مرغ هنوز زنده است.	Morgh hanuz zendeh ast.	pau m ao r g hh ae n uw z z eh n d ax ae s t pau
۳۱- گاو، مادرم را شاخ زد.	Gaav maadaram ra shaakh zad.	pau g aa v m aa d aa r ae m r aa sh aa k z ae d pau
۳۲- اسب، فرار کرد.	Asbe faraar kard.	pau ae s b f aa r aa r k aa d pau
۳۳- رنگ گچ زرد است.	Range gache zard ast.	pau r ae ng g ae ch z aa r d ae s t pau
۳۴- وضع کفش‌هایت خیلی خراب است.	Vazeh kafshhaayat kheili kharaab ast.	pau v ae z ax k ae f sh aa ay ax t k ay l iy k eh r aa b ae s t pau
۳۵- هنوز یخ داریم.	Hanuz yakh daarim.	pau hh ae n uw z y ae k d aa r iy m pau
۳۶- به تو اصلاً ربط ندارد.	Beh toe aslan rabt nadaarad.	pau b eh t ow ae s l ax n r ae b t n aa d aa aa d pau
۳۷- عجب بوی عطری می‌آید.	Ajab booyeh aatari miaayad.	pau ax jh ae b b uw ax ae t r iy m iy aa ay ax d pau
۳۸- قارچ دیر می‌پزد.	Ghaaarch deer mipazad.	pau g aa r k d ih r m ih p ax z ax d pau
۳۹- آب، گرم نیست؟	Ab garm nist?	pau ae b g aa r m n ih s t pau
۴۰- این راه، پایان ندارد.	In raah paayaan nadaarad.	pau ih n r aa p ay aa n n aa d aa aa d pau

```
(phoneme vc vlng vheight vfront vrnd ctype cplace cvox)
(#       -  ۰    -       -      -    ۰     ۰      -)
(a       +  l    ۳       ۱      -    ۰     ۰      -)
(e       +  l    ۲       ۱      -    ۰     ۰      -)
(i       +  l    ۱       ۱      -    ۰     ۰      -)
(o       +  l    ۳       ۳      -    ۰     ۰      -)
(u       +  l    ۱       ۳      +    ۰     ۰      -)
(b       -  ۰    -       -      +    s     l      +)
(ch      -  ۰    -       -      +    a     a      -)
(d       -  ۰    -       -      +    s     a      +)
(f       -  ۰    -       -      +    f     b      -)
(g       -  ۰    -       -      +    s     p      +)
(j       -  ۰    -       -      +    l     a      +)
(k       -  ۰    -       -      +    s     p      -)
(l       -  ۰    -       -      +    l     d      +)
(ll      -  ۰    -       -      +    l     d      +)
(m       -  ۰    -       -      +    n     l      +)
(n       -  ۰    -       -      +    n     d      +)
(ny      -  ۰    -       -      +    n     v      +)
(p       -  ۰    -       -      +    s     l      -)
(r       -  ۰    -       -      +    l     p      +)
(rr      -  ۰    -       -      +    l     p      +)
(s       -  ۰    -       -      +    f     a      +)
(t       -  ۰    -       -      +    s     t      +)
(th      -  ۰    -       -      +    f     d      +)
(x       -  ۰    -       -      +    a     a      -)
  )
)
```

شکل ۳: نمودار درختی تابع Phoneset.

شکل ۳ نمودار درختی تابع Phoneset را نشان می‌دهد. بلوک مربوط به عبارت‌بندی آوایی سخن را قابل‌فهم‌تر می‌کند. انسان بسته به‌اندازه شش‌ها می‌تواند مدت زمان محدودی قبل از نفس‌تازه‌کردن، صحبت کند. این موضوع یک حد بالا برای عبارت‌های آوایی تعیین می‌کند. بااین‌حال، به‌ندرت عبارت‌هایمان را به‌طول بیشینه می‌سازیم. برای زبان انگلیسی (و احتمالاً بسیاری از زبان‌های دیگر) قواعد ساده بر پایه نقطه‌گذاری تعیین‌کننده مناسبی برای حدود عبارت‌های آوایی است. هم‌چنین، این حدود از طریق «ویرگول»، «علامت سؤال»، «علامت تعجب» تعیین می‌شوند. سپس در این بلوک فرکانس گام واج‌های صدادار (واک‌دار) را می‌آوریم.

تغییرات فرکانس گام بر اساس میانگین و انحراف معیار و بزرگ ترین فرکانس گام مربوط هر واج واک‌دار می‌باشد که از طریق محاسبه ارزش مربوط به هر واج با توجه به رابطه ۱ در نرم‌افزار وارد می‌نماییم، برای این منظور به طریق زیر عمل می‌نماییم:

۱- میانگین (mean) و انحراف معیار (stddev) فرکانس‌های گام هر واج واک‌دار را به‌دست می‌آوریم.

۲- بزرگ‌ترین فرکانس گام هر واج واک‌دار را برابر F0 قرار می‌دهیم. در این برنامه مقادیر mm وms بر اساس نرم‌افزار Flite ثابت می‌باشند.

۳- حال از طریق رابطه پیشنهادی ۲ مقدار ارزش را به‌دست می‌آوریم.

$$v = ((((f0 - m)/ s)* ms)+ mm) \quad (۲)$$

که در آن v=value ،m=mean ،s= stddev ،mm=model_mean =۱۷۰ و ms=model_stddev= ۳۴ است.

در رابطه پیشنهادی ۲ با استفاده از سه معیار میانگین، انحراف معیار و بزرگ‌ترین فرکانس گام هر واج واک‌دار و گذاشتن در رابطه پیشنهادی، عددی به‌نام ارزش به‌دست می‌آوریم که این عدد تأثیر مستقیم در طبیعی‌تر و قابل فهم‌تر شدن صدای به‌دست آمده دارد. در واقع مقدار ارزش از طریق مقادیر میانگین، انحراف معیار و بزرگ‌ترین فرکانس گام هر واج واک‌دار تعیین می‌گردد و مقدارش ثابت نمی‌باشد. در برنامه نرم افزار Flite تنها می‌توانیم با تغییر میانگین و انحراف معیار که برای تمامی واج‌ها یکسان تعریف شده است، منحنی فرکانس گام تمام واج‌ها را به مقدار یکسان تغییر دهیم. همچنین مقادیری برای کل واج‌ها ثابت در نظر گرفته شده است که به‌نام ارزش (value)، مدل میانگین (model_mean) و مدل انحراف معیار (model_stddev) نام گذاری شده‌اند. در واقع این سه ثابت هیچ نقشی در فرکانس گام واج‌ها ندارند و برای تمامی واج‌ها یکسان تعریف شده‌اند. بنابراین ما نمی‌توانیم زیروبمی هر واج را جداگانه تغییر دهیم. بنابراین صدای به‌دست آمده بسیار رباتیک می‌باشد و فقط با تغییر عدد مربوط به میانگین و انحراف معیار که برای کلیه واج‌ها می‌باشد، تغییرات یک‌سان به‌کار می‌رود؛ اما مدل پیشنهادی سه خصوصیت مهم دربر دارد:

- برنامه Flite را طوری تغییر دادیم تا بتوانیم زیروبمی هر واج را جداگانه تغییر دهیم. بنابراین برنامه میانگین (mean) و انحراف معیار (stddev) فرکانس‌های گام را برای هر واج واک‌دار جداگانه حساب می نماید.

- در مدل پیشنهادی برای شباهت بیشتر منحنی فرکانس گام به‌دست آمده از روش پیشنهادی با منحنی فرکانس گام پایگاه داده جملات فارسی فارس‌دات و نزدیک‌تر شدن صدای به‌دست آمده از مدل پیشنهادی به صدای انسان از سه معیار میانگین، انحراف معیار و بزرگ‌ترین فرکانس گام هر واج استفاده می‌نماییم. یعنی علاوه بر وارد نمودن میانگین و انحراف معیار فرکانس‌های گام برای هر واج واک‌دار، بزرگ‌ترین فرکانس گام هر واج واک‌دار F_0 را نیز به برنامه اضافه می نماییم.

- همان‌طور که گفتیم در رابطه ۱ مقدار ارزش ثابت در نظر گرفته شده است و با تغییر واج‌ها و تغییر میانگین و انحراف معیار تغییر نمی‌نماید؛ اما با تغییر رابطه ۱ به رابطه پیشنهادی ۲ مقدار ارزش را طوری تعیین نمودیم که حساس به تغییرات سه معیار میانگین، انحراف معیار و بزرگ‌ترین فرکانس گام هر واج باشد و مقدار ارزش برای هر واج جداگانه تعیین می‌گردد.

در این قسمت منحنی گام مربوط به جمله «دیروز برف بارید» که از نرم‌افزار Flite به‌دست آورده‌ایم را با منحنی گامی که از پایگاه داده جملات فارسی فارس‌دات به‌دست آمده است، مقایسه می‌کنیم. همان‌طور که گفته شد، تغییرات فرکانس گام بر اساس میانگین و انحراف معیار فرکانس گام و F0 می‌باشد که از طریق محاسبه ارزش با توجه به رابطه پیشنهادی ۱ به نرم‌افزار وارد می‌شود. در جدول ۴ مقدار میانگین و انحراف معیار فرکانس گام هر واج واک‌دار مربوط به جمله «دیروز برف بارید» را با توجه به پایگاه داده جملات فارسی فارس دات، به‌دست آورده‌ایم. همچنین مقدار بزرگ‌ترین فرکانس گام مربوط به هر واج و مقدار ارزش نشان داده شده است.

در شکل ۴ سیگنال گفتار زمانی و منحنی فرکانس گام جمله «دیروز برف بارید» مربوط به نرم‌افزار Flite بدون اعمال روش پیشنهادی را می‌بینیم و در شکل ۵ سیگنال گفتار زمانی و منحنی فرکانس گام جمله «دیروز برف بارید» مربوط به نرم‌افزار Flite با اعمال روش پیشنهادی را می‌بینیم.

در شکل ۶ منحنی فرکانس گام پایگاه داده جملات فارسی فارس دات، نشان داده شده است. این نمودار را با دسترسی به اطلاعات فرکانس گام جملات فارسی از پایگاه داده فارس‌دات و از طریق نرم‌افزار چایلرز که بر روی برنامه MATLAB ۵/۲ نصب شده است، ترسیم نموده‌ایم.

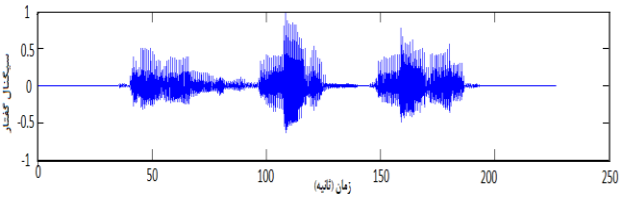
دلیل تفاوت پهنای پالس‌ها در منحنی فرکانس گام «دیروز برف بارید» مربوط به نرم‌افزار Flite با اعمال روش پیشنهادی با منحنی فرکانس گام پایگاه داده جملات فارسی مربوط به جمله «دیروز برف بارید» را می‌توان بدین صورت بیان نمود که در رسم منحنی فرکانس گام پایگاه داده جملات فارسی، تمام فرکانس‌های گام مربوط به هر واج واکدار در ایجاد منحنی دخیل هستند؛ اما در رسم منحنی فرکانس گام

مربوط به نرم‌افزار Flite با اعمال روش پیشنهادی تنها از میانگین و انحراف معیار فرکانس‌های گام هر واج واک‌دار و بزرگ‌ترین فرکانس گام هر واج واک‌دار استفاده می‌نماییم.

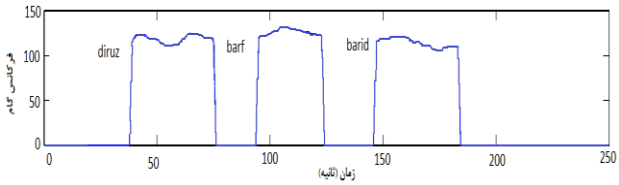
جدول ۴: تعیین میانگین، انحراف معیار، بزرگ‌ترین فرکانس گام و به‌دست آوردن مقدار ارزش مربوط به جمله «دیروز برف بارید».

	d	ih	r	uw	z	b	aa
نوع واج	بی‌واک	واکدار	واکدار	واکدار	واکدار	بی‌واک	واکدار
میانگین	--	۷۷	۱۱۱	۱۳۰	۲۱	--	۶۲
انحراف معیار	--	۵۳	۸	۳	۴۷	--	۵۴
ارزش	--	۱۹۶	۲۲۵	۲۱۵	۲۴۹	--	۲۰۱
بزرگ‌ترین فرکانس گام	--	۱۱۸	۱۲۴	۱۳۴	۱۳۰	--	۱۱۲

	r	f	b	aa	r	ih	d
نوع واج	واکدار	بی‌واک	واکدار	واکدار	واکدار	واکدار	بی‌واک
میانگین	۱۰۵	--	۵۸	۱۷۴	۱۷	۱۰۳	--
انحراف معیار	۳	--	۸۰	۴	۵۱	۷۴	--
ارزش	۲۳۸	--	۲۲۰	۲۱۲	۲۷۳	۲۰۴	--
بزرگ‌ترین فرکانس گام	۱۱۱	--	۱۸۰	۲۲۰	۱۷۲	۱۷۷	--



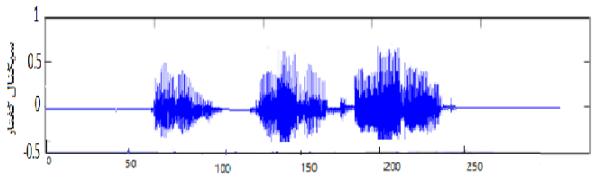
(الف)



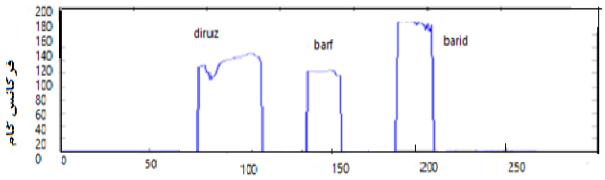
(ب)

شکل ۴: جمله «دیروز برف بارید» بدون اعمال روش پیشنهادی الف) سیگنال گفتار زمانی ب) منحنی فرکانس گام، توسط نرم‌افزار Flite.

همچنین روش پیشنهادی بر روی ۴۰ جمله متفاوت (در جدول ۳ ذکر شده است)، انجام‌گرفته و از مقایسه منحنی فرکانس گام به‌دست آمده از روش پیشنهادی با منحنی فرکانس گام پایگاه داده جملات فارسی فارس دات، نتیجه مطلوب به‌دست آمده است.

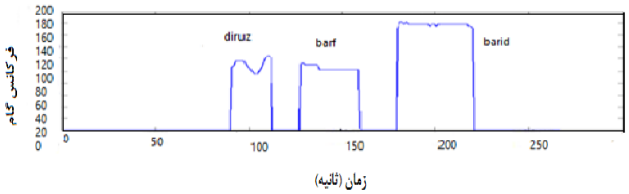


(الف)



(ب)

شکل ۵: جمله «دیروز برف بارید» با توجه به روش پیشنهادی الف) سیگنال گفتار زمانی ب) منحنی فرکانس گام، توسط نرم‌افزار Flite.



شکل ۶: منحنی فرکانس گام پایگاه داده جملات فارسی مربوط به جمله «دیروز برف بارید».

۵- نتایج

برای ارزیابی سیستم‌های سنتز گفتار عمدتاً پارامترهای قابل‌فهم‌بودن، طبیعی‌بودن و خوش‌آیندبودن صدای سنتزشده، مورد ارزیابی قرار می‌گیرد. قابل‌فهم‌بودن، میزان‌وضوح و قابل‌درک‌بودن صدا را نشان می دهد. طبیعی‌بودن، مشخص می‌کند که صدای سنتز شده چقدر به صدای طبیعی انسان نزدیک است و خوش‌آیندبودن نیز مشخص می کند که صدای تولید شده تا چه حدی برای شنونده خوش‌آیند است. برای ارزیابی صدای سنتز شده، روش‌های مختلفی پیشنهاد شده است که آزمون MOS[۴۴] ازجمله مهم‌ترین این روش‌ها به شمار می‌رود [۲۱، ۲۲]. در آزمون MOS، چند جمله مختلف با مشخصات متفاوت از لحاظ گوناگونی واج‌ها، برای تعدادی شنونده پخش می‌شود. هر شنونده یک ارزیابی از پارامترهای قابل‌فهم بودن، طبیعی‌بودن و خوش‌آیندبودن ارائه می‌نماید. این ارزیابی به این صورت است که به‌ازای هرکدام از پارامترهای فوق یک نمره به هر جمله داده می‌شود. نمرات مورد استفاده و مفهوم آن‌ها، در جدول ۵ مشاهده می‌شود.

سپس به‌ازای هرکدام از پارامترهای فوق، از امتیازات میانگین‌گیری می‌کنیم و امتیاز هر پارامتر را مشخص می‌نماییم. همچنین در این پروژه آزمون MOS را بر روی ۴۰ جمله از جملات موجود در پایگاه داده فارسی انجام دادیم که ۲۰ جمله اول مربوط به آموزش نرم‌افزار می‌باشند و ۲۰ جمله دوم برای آزمون نرم‌افزار به‌کار رفته است.

جدول ۵: نمرات و مفهوم آن‌ها در تستMOS .

نمره	مفهوم
۱	بد
۲	ضعیف
۳	نسبتاً خوب
۴	خوب
۵	فوق العاده خوب

برای انجام آزمون MOS، از ۴۰ نفر شنونده استفاده نموده‌ایم. جدول ۶ نتایج حاصل از آزمون MOS، بر روی جملات آموزش با اعمال روش پیشنهادی و بدون اعمال روش پیشنهادی را نشان می‌دهد. همچنین جدول ۷ نتایج حاصل از آزمون MOS، بر روی جملات حاصل از آزمون نرم‌افزار با اعمال روش پیشنهادی و بدون اعمال روش پیشنهادی را نشان می‌دهد.

جدول ۶: مقایسه نتایج حاصل از آزمون MOS، بر روی جملات آموزش با توجه به اعمال روش پیشنهادی و بدون اعمال روش پیشنهادی بر روی نرم‌افزار Flite.

جمله فارسی	قابل فهم بودن		طبیعی بودن		خوشایند بودن	
	روش پیشنهادی	بدون اعمال روش پیشنهادی	روش پیشنهادی	بدون اعمال روش پیشنهادی	روش پیشنهادی	بدون اعمال روش پیشنهادی
۱- دیروز برف بارید.	۳/۹	۳/۱	۳/۸	۳/۲	۴/۸	۳/۶
۲- من یک دستگاه چاپ ساخته‌ام.	۳/۸	۳/۸	۳/۹	۳/۴	۳/۸	۳/۵
۳- پارچ توی کمد است.	۴/۶	۳/۵	۴	۳/۶	۴/۳	۳/۲
۴- کتفم درد می‌کند.	۴/۵	۳/۲	۴/۱	۳/۵	۴/۱	۳/۱
۵- به چه علتی سگ فرار کرد؟	۴/۳	۳/۶	۳/۷	۳/۲	۳/۸	۳/۶
۶- دیروز در کرج، قاسم را دیدم.	۳/۹	۳/۸	۴/۵	۳/۴	۴/۳	۳/۷
۷- گرگ به گله زد.	۳/۷	۳/۳	۴/۴	۳/۳	۴/۵	۳/۵
۸- پایش از مچ شکست.	۴/۳	۳/۷	۳/۸	۳/۲	۴/۷	۳/۶
۹- پاپ فردا می‌آید ایران.	۴/۷	۳/۲	۴/۴	۳/۶	۴/۸	۳/۵
۱۰- سوت گم شد.	۴/۹	۳/۳	۴/۲	۳/۷	۴/۵	۳/۲
۱۱- گاو توی مزرعه است.	۴/۴	۳/۸	۳/۸	۳/۱	۴/۹	۳/۵
۱۲- پیچ باز شد.	۴/۸	۳/۷	۴/۲	۳/۲	۴/۷	۳/۲
۱۳- جیپ راه افتاد.	۳/۷	۳/۲	۴/۶	۳/۷	۴/۹	۳/۵
۱۴- آن اسب چه می خورد؟	۴/۵	۳/۱	۴/۲	۳/۵	۴/۷	۳/۳
۱۵- پمپ، آب ندارد.	۴/۱	۳/۲	۳/۸	۳/۲	۴/۹	۳/۵
۱۶- دو رکعت، نماز خواند.	۴/۵	۳/۴	۴/۵	۳/۷	۴/۷	۳/۵
۱۷- پول را قاب زدم.	۴/۷	۳/۲	۴/۵	۳/۴	۴/۸	۳/۵
۱۸- سرتیپ وقت ندارد.	۴/۴	۳/۳	۴/۳	۳/۸	۴/۶	۳/۲
۱۹- دیروز توت چیدم.	۴/۶	۳/۱	۳/۹	۳/۳	۴/۷	۳/۵
۲۰- کتابت پاره شد.	۴/۹	۳/۲	۴/۷	۳/۲	۴/۵	۳
میانگین	۴/۴	۳/۳	۴/۲	۳/۴	۴/۶	۳/۴

جدول ۷: نتایج حاصل از آزمون MOS، بر روی جملات آزمون با توجه به اعمال روش پیشنهادی و بدون اعمال روش پیشنهادی بر روی نرم‌افزار Flite.

جمله فارسی	قابل فهم بودن		طبیعی بودن		خوشایند بودن	
	روش پیشنهادی	بدون اعمال روش پیشنهادی	روش پیشنهادی	بدون اعمال روش پیشنهادی	روش پیشنهادی	بدون اعمال روش پیشنهادی
۲۱- من لوسش نکرده‌ام.	۳/۹	۳/۳	۴	۳/۶	۴/۱	۳/۷
۲۲- حجاج به مکه رفتند.	۳/۷	۳/۴	۳/۹	۳/۵	۳/۹	۳/۴
۲۳- پاییز فصل برگ ریزان است.	۴/۱	۳/۲	۳/۹	۳/۴	۴/۸	۳
۲۴- داس، شکست.	۴/۳	۳/۲	۴/۲	۳/۵	۴/۳	۳/۲
۲۵- پارچ لب طاقچه است.	۴/۶	۳/۶	۴/۵	۳/۳	۴/۴	۳/۳
۲۶- سگ علی را گاز گرفت.	۴/۵	۳/۳	۳/۶	۳/۲	۴/۶	۳/۴
۲۷- این ظرف، در ندارد؟	۴/۲	۳/۵	۴/۲	۳/۵	۳/۸	۳/۵
۲۸- فقر باعث یاس می‌شود.	۴/۴	۳/۲	۴/۴	۳/۴	۳/۷	۳/۱
۲۹- جنب خانه ما یک بانک است.	۳/۹	۲/۸	۴/۶	۳/۲	۳/۹	۳
۳۰- مرغ هنوز زنده است.	۴/۴	۳/۲	۳/۷	۳/۳	۴/۵	۳/۵
۳۱- گاو، مادرم را شاخ زد.	۳/۹	۳/۲	۳/۸	۳/۵	۳/۸	۳/۴
۳۲- اسب، فرار کرد.	۳/۹	۳/۳	۴/۵	۳/۶	۴/۹	۳/۱
۳۳- رنگ گچ زرد است.	۳/۷	۳/۴	۴/۳	۳/۶	۴/۲	۳/۴
۳۴- وضع کفش هایت خیلی خراب است.	۴/۱	۳/۲	۳/۶	۳/۱	۴/۶	۳/۶
۳۵- هنوز یخ داریم.	۴/۷	۳/۲	۳/۸	۳/۴	۳/۹	۳/۱
۳۶- به تو اصلاً ربط ندارد.	۳/۸	۳/۴	۴/۶	۳	۴/۴	۳/۷
۳۷- عجب بوی عطری می‌آید.	۴/۳	۳/۳	۳/۹	۳/۳	۴/۵	۳/۲
۳۸- قارچ دیر می پزد.	۴/۴	۳/۴	۴	۳/۵	۳/۸	۳/۴
۳۹- آب، گرم نیست؟	۳/۹	۳/۴	۳/۹	۳/۲	۳/۷	۳/۲
۴۰- این راه، پایان ندارد.	۴/۵	۳/۶	۴/۶	۳/۲	۴/۸	۳/۵
میانگین	۴/۲	۳/۳۹	۴/۱	۳/۳۶۵	۴/۳	۳/۳۳۵

جدول ۸: مقایسه نتایج حاصل از آزمون MOS بر روی سنتز جملات آزمون روش پیشنهادی با روش‌های مشابه پیشین.

	قابل فهم بودن	طبیعی بودن	خوشایند بودن
جملات سنتز شده با استفاده از روش پیشنهادی اعمال شده بر Fliteروی نرم افزار	۴/۴	۴/۲	۴/۸
جملات سنتز شده بدون اعمال روش پیشنهادی Fliteروی نرم افزار	۳/۳	۳/۴	۳/۴
جملات سنتز شده با استفاده از مدلهای ترایفون برای طول واج و گام [۶]	۳/۸	۳/۹	۳/۵
جملات سنتز شده با استفاده از درخت تصمیم گیری برای طول زمانی واج و گام [۶]	۴/۲	۴/۴	۴/۱
جملات سنتز شده با استفاده از درخت تصمیم گیری برای طول زمانی واج‌ها و روش اتوماتیک برای تولید گام [۶]	۴/۶	۴/۳	۴/۵
سنتز کننده پیاده سازی شده به روش هارمونیک نویزی [۷]	۴/۱۸	۳/۴۲	۳

نتایج به‌دست‌آمده توسط آزمون MOS، برای پارامترهای قابل‌فهم‌بودن، طبیعی‌بودن و خوش‌آیندبودن برای جملات آموزش به‌ترتیب ۴/۴، ۴/۲ و ۴/۶ می‌باشند. هم‌چنین برای جملات جهت آزمون نرم‌افزار، پارامترهای قابل‌فهم‌بودن، طبیعی‌بودن و خوش‌آیندبودن به‌ترتیب برابر ۴/۲، ۴/۱ و ۴/۳ می‌باشند. مقایسه این نتایج با حاصل روش‌های مشابه پیشین در جدول ۸ آورده شده است.

۶- نتیجه‌گیری

در این پژوهش، سنتز گفتار به زبان فارسی با استفاده از نرم‌افزار Flite مورد بررسی قرار گرفت. تغییرات منحنی گام در نرم‌افزار Flite با استفاده از تغییر میانگین و انحراف معیار فرکانس گام می‌باشد که دارای دو مشکل است:

برای تمام واج‌ها، یک میانگین و انحراف معیار واحد تعریف شده است یعنی با تغییر میانگین و انحراف معیار تعریف‌شده، منحنی گام تمام واج‌ها به‌یک‌اندازه تغییر می‌کند. همچنین در این نرم‌افزار برای صداهای بی‌واک نیز فرکانس گام در نظر گرفته شده است. بدین‌طریق که به‌وسیله درون‌یابی صداهای واک‌دار مجاور، به صدای بی‌واک (غیر از سکوت) فرکانس گام اختصاص داده و نتیجه آن هموارسازی می‌شود؛ بنابراین تشخیص صداهای واک‌دار از بی‌واک وجود ندارد.

در مدل پیشنهادی ابتدا با توجه به پایگاه داده جملات فارسی فارس‌دات، میانگین، انحراف معیار و بزرگ‌ترین فرکانس گام هر واج واک‌دار را به‌دست می‌آوریم، سپس تغییرات هر واج عبارت موردنظر را از طریق محاسبه مقدار ارزش آن با توجه به رابطه ارائه‌شده در بخش روش پیشنهادی، در نرم‌افزار وارد می‌کنیم تا مشکل اول حل شود و منحنی گام هر واج با توجه به مقدار میانگین، انحراف معیار و بزرگ ترین فرکانس گام خودش تغییر کند، هم‌چنین در این روش فقط برای واج‌های واک‌دار فرکانس گام در نظر گرفتیم و برای واج‌های بی‌واک، میانگین و انحراف معیاری در نظر گرفته نشد تا واج‌های واک‌دار از بی‌واک تشخیص داده شوند و مشکل دوم نیز حل شود. نتایج حاصل از آزمون‌های شنیداری، برای میزان قابل‌فهم بودن، طبیعی بودن و خوش آیندبودن برای جملات آموزش به‌ترتیب ۴/۴، ۴/۲ و ۴/۶ می‌باشد. همچنین، نتایج برای جملات مجموعه آزمون، به‌ترتیب برابر با ۴/۲، ۴/۱ و ۴/۳ می‌باشد.

مراجع

[1] A. B. Black and K. A. Lenzo, *Building synthetic voices*, For FestVox 2.1 Edition, 2007.

[۲] محمدمهدی همایون پور، محمد ایزدی، «تبدیل حرف به صدا در سیستم‌های تبدیل متن به گفتار فارسی با استفاده از درخت‌های تصمیم‌گیری CART»، *دوازدهمین کنفرانس سالانه انجمن کامپیوتر ایران*، تهران، ۱۳۸۵.

[3] R. Kurzweil, *The singularity is near*, Penguin Books, ISBN 9-303788-14-0, 2005.

[4] A.B. Black and K.A. Lenzo, *Flite: a small, fast speech synthesis engine*, System documentation Edition 1.3, for Flite version 1.3, 2005.

[5] A. B. Black, P. Taylor and R. Caley, *The Festival Speech Synthesis System*, [Online], Available: http://www.cstr.ed.ac.uk/projects/festival.-html, 1998.

[۶] محمدمهدی همایون پور، مجید نم نبات، «تبدیل حرف به صدا در زبان فارسی به کمک شبکه‌های عصبی پرسپترون چندلایه‌ای»، *فصلنامه مهندسی برق و مهندسی کامپیوتر ایران*، شماره ۳، صفحات ۱۵۴-۱۴۷، پائیز ۱۳۸۶.

[۷] محمدمهدی همایون پور، سیدمصطفی موسوی، «تولید پارامترهای سنتز گفتار فارسی با استفاده از مدل‌های مخفی مارکوف و درخت تصمیم‌گیری»، *نشریه علمی - پژوهشی انجمن کامپیوتر ایران*، شماره ۱ و ۳ (الف)، صفحات ۱۹-۳۰، بهار و پائیز ۱۳۸۳.

[8] Y. Sagisaka, "Speech synthesis from text," *IEEE Commun. Mag.*, pp. 35-41, 1990.

[۹] منصور شیخان نصیرزاده. مجید و دفتریان. علی، «طراحی و پیاده‌سازی سیستم تبدیل متن به گفتار طبیعی برای زبان فارسی»، *مجله علمی- پژوهشی دانشکده مهندسی دانشگاه فردوسی مشهد*، شماره ۲، صفحات ۳۱-۴۸، ۱۳۸۴.

[۱۰] محمدمهدی همایون پور، آرمین سلیمی بدر، «تعیین مرز و نوع عبارات نحوی در متون فارسی»، *فصلنامه علمی - پژوهشی پردازش علائم و داده‌ها*، شماره ۲، صفحات ۶۹-۸۶، ۱۳۹۲.

[11] N. Thorensen, "Sentence intonation in textual context-supplementary data," *J. Acoust. Soc. Am.*, vol. 80, no. 4, pp. 1041-1047, 1986.

[12] Y. Sagisaka, "On the prediction of global F_0 shape for Japanese TTS," *In Proc. ICASSP*, USA, pp. 325-328, 1990.
[13] J. Buhmann, et al. "Intonation modeling for the synthesis of structured documents," *In Proc. ICSLP*, USA, pp. 2089-2092, 2002.
[14] M. Riedi, "A neural-network-based model of segmental duration for speech synthesis," *In Proc. Eurospeech*, Spain, pp. 599-602, 1995.
[15] Z. Yiqing, "Syllable duration and its functions in standard Chinese discourse," *In Proc. ICSLP*, p. 1097, China, 2000.
[16] C.L. Smith, "Modeling durational variability in reading aloud a connected text," *In Proc. ICSLP*, pp. 1769-1772, USA, 2002.
[17] Y. Sagisaka, and Sato H. "Accentuation rules in Japanese TTS conversion," *Rev. Elect. Commun. Lab.*, vol. 32, no. 2, pp. 188-199, 1984.
[18] P. H. Low and S. Vaseghi, "Application of microprosody models in TTS synthesis," *In Proc. ICSLP*, pp. 2413-2416, USA, 2002.
[19] Y. Hifny. and M. Rashwan, "Duration modeling for Arabic TTS synthesis," *In Proc. ICSLP*, pp. 1773-1776, 2002.
[20] A. Breen, "Speech synthesis models: a review," *Elect. Commun. Engng. J.*, pp. 19-31, 1992.
[۲۱] سیدسعید آیت، *طراحی و پیاده‌سازی سیستم تولید گفتار فارسی با تأکید بر بهبود هرچه بیشتر کیفیت گفتار تولیدشده*، پایان‌نامه کارشناسی ارشد، دانشگاه صنعتی امیرکبیر، ۱۳۷۹.
[22] S. Lemmetty, *Review of Speech Synthesis Technology*, Master Thesis, Helsinki University of Technology, 1999.

زیرنویس‌ها

[1] FarsDat: (Farsi Speech Database) www.dadegan.ir
[2] Stddev (Standard deviation)
[3] Phoneme
[4] Value
[5] Source- filter
[6] Concatenative
[7] LPC (Linear predictive coding)
[۸] TTS (Text to speech)
[9] NLP (Natural Language Processor)
[10] Morphological
[11] POS (Part-Of-Speech tagging)
[12] Prosody
[13] Voiced
[14] Unvoiced
[15] Duration
[16] Intensity
[17] Pause
[18] Articulatory Synthesis
[19] Formant Synthesis
[20] Concatenative Synthesis
[21] Vocal tract
[22] Time-Domain Pitch Synchronous OverLap Add
[23] Pitch marks
[24] Residual
[25] Vowel
[26] Consonant
[27] Vowel or Consonant
[28] Diphthong
[29] Schwa
[30] Vowel length
[31] Vowel frontless
[32] Lip rounding
[33] Fricative
[34] Affricative
[35] Nasal
[36] Semi_vowel liquid
[37] Vonsonant type
[38] Labial
[39] Alveolar
[40] Palatal
[41] Velar
[42] Place of articulation
[43] Consonant or Voicing
[44] Mean opinion score