نشریه پژوهش‌های زبان‌شناسی تطبیقی
سال ششم- شماره ۱۲- پاییز و زمستان ۱۳۹۵

طراحی و ضبط پایگاه‌دادگان گفتاری برای سیستم‌های تبدیل متن به گفتار فارسی*

مرتضی طاهری اردلی[۱*]
سهیل خرم[۲]
مصطفی عاصی[۳]
حسین صامتی[۴]
محمود بی‌جن‌خان[۴]

چکیده

سیستم‌های تبدیل متن به گفتار نرم‌افزارهای رایانه‌ای هستند که می‌توانند برای کمک به افراد کم‌بینا و نابینا مورد استفاده قرار گیرند. برای ساخت این سیستم‌ها و دستیابی به گفتار مطلوب، نیاز به طراحی و آماده‌سازی یک پایگاه‌دادگان گفتاری مناسب است. مقالهٔ حاضر به ارائهٔ روشی برای طراحی و ساختِ پایگاه‌دادگانی مختص سیستم‌های تبدیل متن به گفتار با در نظر گرفتن ساخت نوایی فارسی می‌پردازد. این مجموعه به لحاظ آوایی و نوایی غنی و مشتمل بر ۲۸۲۶ نمونه جملهٔ فارسی است. این نمونه جملات در شرایط استودیو و محیطی عاری از نوفه و با تک‌صدای گویندهٔ خانم که به صورت حرفه‌ای در زمینهٔ صدا فعالیت می‌کند ضبط شده است. پوشش حالت‌های مختلف نوایی در کنار پوشش حالت‌های مختلف آوایی از نقاط قوت این پایگاه است که برای نخستین بار در سیستم‌های تبدیل متن به گفتارِ فارسی لحاظ می‌شود. این مجموعه در کنار فایل‌های صوتی، دارای برچسب‌های متن و صورت آوایی است که به صورت دستی اصلاح شده‌اند. در نهایت، با بکارگیری مجموعه دادگان مذکور و با استفاده از روش بازسازی گفتار آماری ـ پارامتری ساخت صدا انجام گرفت. آزمودنی‌ها کیفیت صدای ساخته‌شده را با استفاده از معیار میانگینِ امتیازاتِ نظردهی (MOS) ۴٫۳ ارزیابی کردند.

کلیدواژه‌ها: پایگاه‌دادگان گفتاری، تبدیل متن به گفتار، نوای گفتار، پیکرهٔ متنی.

* این مقاله برگرفته از رسالهٔ دکتری نگارنده است.

۱- استادیار دانشگاه شهرکرد*(نویسنده مسئول) ✉| taheriling@gmail.com
۲- دانش آموخته دکتری دانشگاه صنعتی شریف ✉| khorram@ce.sharif.edu
۳- استاد پژوهشگاه علوم انسانی و مطالعات فرهنگی ✉| s_m_assi@ihcs.ac.ir
۴- دانشیار دانشگاه صنعتی شریف ✉| sameti@sharif.edu
۵- دانشیار دانشگاه تهران ✉| mbjkhan@ut.ac.ir

تاریخ دریافت: ۹۴/۰۵/۰۷ تاریخ پذیرش: ۹۵/۰۹/۲۴

مقدمه

گفتار اصلی‌ترین راه ارتباطی انسان‌ها با یکدیگر است و استفاده از آن می‌تواند ارتباط بین انسان و ماشین را آسان‌تر کند. این ارتباط از طریق دو بخش پردازش گفتاری با نام‌های بازشناسی گفتار (speech recognition) و تبدیل متن به گفتار (text-to-speech) میسر می‌شود که در کنار ترجمهٔ ماشینی (machine translation)، مهمترین کاربردهایِ فناوریِ پردازش زبان طبیعی (natural language processing) محسوب می‌شوند (ژورافسکی و مارتین، ۲۰۰۷). توسعه و پیاده‌سازی هر یک از این حوزه‌ها مستلزم منابع گوناگون از جمله پیکره‌های متنی و پایگاه‌دادگان‌های گفتاری[۱] است که عموماً زبان‌ویژه‌اند. در ارتباطِ گفتاری بین ماشین و انسان، سیستم تبدیل متن به گفتار نقش خروجی را ایفا می‌کند که این خروجی، گفتار بازسازی‌شدهٔ انسان است. به‌طور کلی، سیستم‌های تبدیل متن به گفتار که توانایی خواندن متون یک یا چندین زبان را دارند از دو بخش پایه‌ای تشکیل شده‌اند؛ بخش نخست، وظیفهٔ استخراج اطلاعات آوایی و نوایی از متن را برعهده دارد و بخش دوم که بازسازی گفتار (speech synthesis) نامیده می‌شود در تبدیل اطلاعات آوایی و نوایی به موج گفتار (speech wave) کاربرد دارد. در حقیقت، تبدیل متن به گفتار تلاش برای تقلید توانایی‌های انسان در خواندن متون است که این متون خود ممکن است از طریق صفحه کلید و یا به صورت یک فایل متنی و یا پس از شناسایی توسط یک سیستم نویسه‌خوان نوری (Optical Character Recognition (OCR)) دریافت شوند (همایون‌پور، ۱۳۹۱: ۲۱).

تیلور[۲] (۲۰۰۹: ۴۱) شیوهٔ عملکردِ غالبِ سیستم‌های مذکور را این گونه ترسیم می‌کند که در ابتدا، متن به صورت توالی‌ای از جملات به سیستم ارائه می‌شود. سپس، به منظور مدیریت هرچه بهتر فرایند، به کمک الگوریتم مشخصی، مرز جملات تعیین می‌شود تا فرایند صداسازی، جمله به جمله انجام گیرد. در این بخش، با معیار قرار دادن فاصلهٔ سفید (white space) و علائم نقطه‌گذاری، جملهٔ مورد نظر به قطعه‌ها (token) تقسیم می‌شود که این قطعه‌ها خود ممکن است شامل رمزگانِ (encoding) مختلفی مانند کلمات، اعداد، تاریخ و دیگر اطلاعات باشند. در این مرحله، رمزگان‌های مختلف یکدست می‌شوند و به نشانه‌های زبان

---

۱. در متون مختلف، دو اصطلاح پایگاه‌دادگان (database) و پیکره (corpus) به صورت جایگزین به جای هم به‌کار رفته‌اند. کمپل (۲۰۰۵) تمایز و تفاوت این دو را اینگونه مطرح می‌کند که «پایگاه‌دادگان» مجموعه‌ای نظام‌مند از قطعه‌هاست که به صورت کنترل‌شده و هدفمند *طراحی* شده است در حالی‌که پیکره مجموعه‌ای از نمونه‌هاست که از مطالب موجود و بدون ترتیب خاصی *جمع‌آوری* می‌شوند و می‌توان برپایهٔ آن پایگاه‌دادگان خاصی طراحی کرد. از جمله تفاوت دیگر این دو، میزان حجم و اندازه است؛ بر این مبنا که پایگاه‌دادگان دارای حجمی محدود است در حالی که پیکره باید به اندازه کافی بزرگ باشد تا بتواند نمایندهٔ واقعیتِ رخدادها و الگوهای ممکن زبانی باشد. حال نکته اینجاست که اصطلاح بازسازی گفتارِ پیکره‌بنیاد (corpus-based speech synthesis) یک اصطلاح جاافتاده و مرسوم است که تمایز فوق را با مشکل مواجه می‌کند. از طرفی دیگر، در متون فارسی، اصطلاح پیکره، دادگان، پایگاه‌داده و پایگاه‌دادگان در جایگاه‌های مختلفی مرتبط با موضوع فوق به‌کار رفته‌اند که مشکل انتخاب اصطلاح مناسب را دوچندان می‌کند. در تحقیق حاضر، برپایهٔ تمایز کمپل (۲۰۰۵)، «پایگاه‌دادگان» برای مجموعهٔ نظام‌مندِ طراحی‌شده از داده‌ها و اصطلاح پیکره برای مجموعه‌ای جمع‌آوری‌شده از داده‌ها به‌کار رفته است.

2. Taylor

طبیعی یعنی کلمات ترجمه می‌شوند تا در مرحلهٔ بعد صورت آوایی (phonetic form) آنها مشخص شود. سپس سعی می‌شود تا با اطلاعات موجود یعنی متن و صورت آوایی تحلیل نوایی (prosodic) انجام گیرد. اگر چه در متن اطلاعات ناچیزی پیرامون نوای گفتار وجود دارد اما تمام تلاش‌ها صورت می‌گیرد تا از الگوریتم‌هایی استفاده شود تا گروه‌بندی (phrasing)، الگوهای برجستگی و ساخت آهنگی استخراج شود. تا این مرحله، مجموعه فرایندهای استخراج اطلاعات آوایی و نوایی از متن ورودی که تحلیل متن و نوای گفتار نامیده می‌شود انجام گرفته است. اما در بخش دوم، یعنی بازسازی گفتار، کلمات، آواها و گروه‌بندی نوایی، نقش ورودی را برای انتخاب واحدها از یک مجموعهٔ از پیش ضبط‌شده که پایگاه‌دادگان گفتاری نامیده می‌شود ایفا می‌کنند. در این مسیر، بهینه‌ترین واحدها از بین مجموعه واحدهای آوایی و نوایی انتخاب می‌شود تا در کنار یکدیگر قرار گیرند و در نهایت سیگنال گفتاری تولید شود. اما از آنجایی که نمی‌توان تمام کلمات و جملات ممکن زبان را در یک مجموعه فراهم کرد، از مهمترین بخش‌ها در یک سیستم، شیوهٔ آماده‌سازی و تهیهٔ پایگاه‌دادگان گفتاری است تا بتوان مناسبترین واحدها را در آن لحاظ کرد. به بیان دیگر، کیفیت خروجی یک سیستم تبدیل متن به گفتار به شدت وابسته به چگونگی طراحی پایگاه‌دادگان است به گونه‌ای که هر خطای احتمالی در آن، در صدای بازسازی‌شده منعکس می‌شود. از آنجایی که هدف اصلی سیستم‌های تبدیل متن به گفتار دستیابی به صدایِ بازسازی‌شدهٔ قابل‌فهم (intelligible) و طبیعی (natural) است (تیلور، ۲۰۰۹: ۵۲۳)، به منظور دستیابی به صدایی طبیعی‌تر، تنها نمی‌توان به ضبط یک نمونه از یک واحد مانند هجا یا دوآوایی (diphone) بسنده کرد، بلکه انتخاب نمونه‌هایی از یک هجای مشخص در جایگاه‌های نوایی مختلف از اهمیت ویژه‌ای برخوردار است که در این مقاله به آن خواهیم پرداخت.

از پایگاه‌دادگان‌های مختص بازسازی گفتار که تاکنون در زبان فارسی گزارش شده است، می‌توان به اسلامی و همکاران (۱۳۸۸) و آیت (۱۳۸۹) اشاره کرد. اسلامی و همکاران (۱۳۸۸) به طراحی و ساخت دو پایگاه‌دادگان هجایی و دوآوایی (دایفونی) پرداخته‌اند. در پایگاه‌دادگان هجایی، با حائز اهمیت خواندن جایگاهِ هجاها در زنجیرهٔ گفتار و پس از انجام آزمون شنیداری، جایگاه ابتدایی کلمه، مناسبترین جایگاه برای ضبط انواع هجاهای ممکن (حدود شش هزار هجا) در نظر گرفته شد. برای تهیهٔ پایگاه‌دادگان دوآوایی نیز با در نظر گرفتن اصول مشخصی مانند استخراج دوآوایی‌ها از جایگاه فاقد تکیه و همچنین عدم استفاده از کلمات تک‌هجایی، مجموعه‌ای بالغ بر ۹۶۴ دوآواییِ فارسی در تک‌کلمه‌ها و کلمات مرکب، طراحی، ضبط و در نهایت تقطیع شد. در اثری دیگر، آیت (۱۳۸۹) به طراحی پایگاه‌دادگان دوآواییِ فارسی پرداخت. برای رسیدن به این هدف، بعد از ضبط کلمات و جملات حاوی دوآوایی‌هایِ ممکن فارسی، عمل تقطیع به کمک روش شنیداری، بررسی سیگنال زمانی و نیز طیف‌نگاشت انجام شد.

مقالهٔ حاضر از بخش‌های مختلفی تشکیل شده است. در ابتدا، نگاهی کلی به سیستم‌های بازسازی‌کنندهٔ گفتار خواهیم داشت. سپس، به چگونگی پوشش آوایی و نوایی می‌پردازیم. بعد از چگونگی پوشش آوایی، به طراحی و انتخاب نمونه جملات (prompt set) و انتخاب گوینده پرداخته شده است. در نهایت، ضبط صدا و

بررسی فایل‌های صوتی، میزان پوشش آوایی و نوایی و ساخت صدا با استفاده از پایگاه طراحی‌شده و ارزیابی حاصل از آن اشاره شده است.

سیستم‌های بازسازی گفتار

امروزه با افزایش توان و منابع محاسباتی در فناوری رایانه، ساخت صدا از یک رویکرد دانش‌بنیاد (knowledge-based) به یک رویکرد داده‌بنیاد (data-based) بدل شده است. از آغاز بکارگیری پایگاه‌دادگان‌هایی از گفتار طبیعی، تکامل سیستم‌های بازسازی گفتار را می‌توان از سیستم‌های دوآوایی (مولینه و شارپنتیه، ۱۹۹۰) تا سیستم‌های با کاربرد عام‌تر مانند انتخاب واحد (Unit Selection) (هانت و بِلک، ۱۹۹۶) در نظر گرفت که در آن واحدها به صورت خودکار از پایگاه‌دادگانِ گفتاری مورد نظر انتخاب می‌شوند (زِن و همکاران، ۲۰۰۹: ۱). در دهۀ پایانی قرن بیستم اکثر سیستم‌ها با استفاده از این رویکرد یعنی انتخاب واحد ساخته شده‌اند که نیازمند حجم زیادی از داده‌های گفتاری است. در نقطۀ مقابلِ روش‌هایی از نوع هم‌گذاری (concatenative) مانند انتخاب واحد، روشِ بازسازیِ گفتارِ آماری ـ پارامتری (Statistical Parametric) (یوشیمورا و همکاران، ۱۹۹۹؛ لینگ و همکاران، ۲۰۰۶؛ بلک، ۲۰۰۶؛ زِن و همکاران، ۲۰۰۷) قرار دارد که در سال‌های اخیر با اقبال خوبی روبرو شده است و در حقیقت، بدل به یک رقیب جدی برای سیستم‌هایی مانند انتخاب واحد شده است (زِن و همکاران، ۲۰۰۹: ۱). از نمونه‌های بارز روشِ آماری ـ پارامتری می‌توان به بازسازی گفتار مبتنی بر مدل مخفی مارکف ((Hidden Markov Model (HMM) اشاره کرد. برخلاف رویکرد پیشین که واحدهای زنجیرۀ گفتار را بدون تغییر انتخاب می‌کند، در این روش فنی یعنی آماری ـ پارامتری، میانگین مجموعۀ مشخصی از عناصر مشابه از پایگاه‌دادگان را برای صدای خروجی لحاظ می‌کند[1] (زِن و همکاران، ۲۰۰۹: ۱). پایگاه‌دادگان حاضر برای اتخاذ رویکردی آماری–پارامتری طراحی شده است که در مقایسه با سیستم رقیب، نیازمند حجم کمتری از صداهای از پیش ضبط‌شده است. حوزۀ کاربردی این پایگاه‌دادگان به منظور استفاده در سامانۀ خبرخوان فارسی است که انتظار می‌رود قابلیت خواندنِ مطلوب هر نوع متن جدید در این سیاق را داشته باشد.

پوشش واحدهای آوایی و نوایی

با نگاهی اجمالی به سیستم‌های تبدیل متن به گفتار، چنین به نظر می‌رسد که ابتدا باید متن ورودی به شکل دنبالۀ آوایی متناظر با آن تبدیل و سپس این زنجیرۀ آوایی بیان شود. اما انسان برای انتقال مفاهیم تنها از بیان پی‌درپی آواها استفاده نمی‌کند بلکه از ساختار زبرزنجیری (suprasegmental) مانند آهنگ در سطح جمله نیز استفاده می‌کند. اطلاعات آهنگِ گفتار نقش برجسته و ممتازی در رساندن معنیِ پیام بازی می‌کنند تا آنجا که اطلاعاتی که از طریق نحوۀ بیان جمله یا عبارت منتقل می‌شود به مراتب بیشتر از اطلاعاتی است که در معنای واژگانی کلمات وجود دارد (هویزینگر، ۱۹۹۹: ۱۳). از طرفی، کیفیت صدای خروجی سیستم‌های

۱. به منظور کسب اطلاعات بیشتر پیرامون رویکرد مدل مخفی مارکف (HMM) رجوع شود به طاهری اردلی و خرم (۱۳۹۱).

تبدیل متن به گفتار به‌شدت وابسته به واحدهای آکوستیکیِ (acoustic unit) پوشش داده شده دارد. عوامل متعددی در کیفیت این واحدها دخیل هستند که از آن جمله می‌توان به نوع واحدهای انتخابی (تک‌آوا، دوآوا، سه‌آوا (triphone) هجا و غیره)، دقت در برچسب‌گذاری (labeling)، تعداد نمونه‌ها از هر واحد و همچنین غنی‌بودنِ واحدها به لحاظ نوایی اشاره کرد (ماتوشک و همکاران، ۲۰۰۸). اما همان‌طور که در بخش مقدمه اشاره شد لحاظ کردن کلیهٔ جملات و عبارت‌های ممکن زبان در یک مجموعه امکان‌پذیر نیست. بنابراین، به منظور طراحی و ساخت یک پایگاه‌دادگان غنی، نه‌تنها باید پوشش واحدهایی زنجیری مانند انواع هجاها مدنظر قرار گیرد بلکه حالت‌های مختلف نوایی آنها نیز باید لحاظ شود؛ به بیان دیگر، ضبط یک هجای مشخص در جایگاه‌های نوایی مختلف. انتخاب هجا به عنوان واحد اصلی مورد استفاده برای پوشش واحدهای آکوستیکی به جهت اهمیت آن به عنوان کوچکترین واحد در سلسله‌مراتب نوایی (prosodic hierarchy) است (نسپور و فوگل، ۲۰۰۷). درواقع، هجا کوچکترین واحدی است که از آن می‌توان برای پوششِ نه تنها واحدهای آکوستیکیِ آوایی (زنجیری) بلکه واحدهای آکوستیکیِ نوایی (زبرزنجیری) نیز بهره برد. در ضمن، ساختمان سادهٔ هجا و نیز سهولت در تشخیص مرز آن در فارسی، دلیل دیگر در انتخاب هجا به عنوان واحد اصلی مورد استفاده در پایگاه‌دادگان حاضر است (اسلامی و همکاران، ۱۳۸۸). از طرفی دیگر، از چشم‌انداز نوای گفتار، برجستگی یک هجا در سطح کلمه در فارسی از نوع زیروبمی (pitch) است و ماهیتاً با آنچه که در زبان‌های ژرمنیِ غربی مانند انگلیسی و هلندی از آن تحت عنوان تکیهٔ واژگانی (lexical stress) یاد می‌کنند متفاوت است (ابوالحسنی‌زاده و همکاران، ۲۰۱۲). به بیان ساده‌تر، از بین مؤلفه‌های بسامد پایه (fundamental frequency)، دیرش (duration) و شدت (intensity)، بسامدپایه مهمترین عامل تعیین‌کنندهٔ نوای گفتار فارسی است (ابوالحسنی‌زاده و همکاران، ۲۰۱۲؛ طاهری اردلی و ژو، ۲۰۱۲). بنابراین، در نظر گرفتن ساخت‌های متفاوت زیروبمی باید در دستور کار قرار گیرد. حال سؤال اصلی اینجاست که به چه شکل باید چنین ساخت‌هایی را در یک پایگاه‌دادگان لحاظ کرد؟ به بیان دیگر، چگونه می‌توان چنین حالت‌هایی را در نمونه جملاتِ طراحی‌شده در نظر گرفت تا گویندهٔ حرفه‌ای رخدادهای آهنگی مورد نظر را تولید کند. از این رو، سعی شد تا هجاهای مختلف در جایگاه‌های گوناگون جمله، یعنی جایگاه‌هایی که آشکارا بیشترین تفاوت‌های نوایی محرز است، ضبط شوند. این جایگاه‌ها عبارتند از: ابتدای جمله، قبل از نقطه (پایان جمله)، قبل از کاما، قبل از دونقطه و در سایر جایگاه‌ها (غیر از جایگاه‌های مذکور). اما انتخاب این جایگاه‌ها از آن جهت است که انتخاب رخدادهای آهنگی متفاوت را از روی پیکرهٔ متنی ممکن می‌سازد و از طرف دیگر، تولید یک رخداد آهنگی مشخص را توسط گوینده تضمین می‌کند. در چرایی استفاده از این جایگاه‌ها باید گفت که هجای آغازی جملات، غالباً دارای آهنگ متفاوتی است و با نواخت بالا (high) تولید می‌شود (سادات تهرانی، ۲۰۰۷). هجای انتهایی جملات یعنی قبل از نقطه، نشانگر تمام شدن عبارت است از این رو حامل نواخت مرزنمای (boundary tone) خبری است که به صورت پایین (low) تولید می‌شود. جایگاه قبل از کاما، نشانگر آهنگ غیرپایانی (non-final) است که به لحاظ واجی متفاوت از دیگر رخدادهای نواختی عمل می‌کند. انتخاب جایگاه قبل از دونقطه از آن جهت است که سیستم نهایی قرار است به عنوان

یک خبرخوان مورد استفاده قرار گیرد و در این نوع سیستم‌ها کاربرد نقل‌قول‌های مستقیم به‌وفور یافت می‌شود. همچنین، به منظور اهمیت جایگاه تکیه، هجای دارای هستهٔ تکیه‌بر به‌عنوان یک واحد مجزا در نظر گرفته شد. هجای تکیه‌بر در فارسی به صورت بالقوه می‌تواند جایگاه نواختِ بالا در یک تکیهٔ زیروبمی (pitch accent) باشد. در ضمن، تلاش شد تا کلیهٔ دوآوایی‌های ممکن فارسی نیز در پایگاه‌دادگان لحاظ شود. پوشش دوآوایی‌ها فارغ از جایگاه آنها در بافت نوایی است.

طراحی و انتخاب خودکارِ نمونه جملات

یکی از دشوارترین مراحلِ ساخت پایگاه‌دادگان، چگونگی طراحی و انتخابِ نمونه جملات است. در پژوهش حاضر، این مرحله را می‌توان به دو بخش تقسیم کرد: نخست استفاده از نمونه جملاتِ پایگاه‌دادگانی که پیش از این در یک سیستم تبدیل متن به گفتارِ فارسی استفاده شد، سیستمی که در واقع بیشتر بر روی پوشش واحدهای دوآوایی تمرکز دارد (خرم و همکاران، ۲۰۱۴). این مجموعه که شامل ۱۳۰۰ نمونه جمله است به معیارهای جدید ویرایش شد و در نهایت ۴۵۶ مورد از ۱۳۰۰ نمونه جمله حذف شدند. اما بخش عمدهٔ نمونه جملات مستقیماً از یک پیکرهٔ متنی ده‌میلیون‌کلمه‌ای استخراج شد[۱]. در این مسیر، در ابتدا به کمک الگوریتمی عملیات جداسازی قطعه‌ها انجام گرفت. سپس فهرستی از بسامد کلمات و هجاهای موجود در پیکره بدست آمد و جملات خبری، پرسشی و تعجبی از یکدیگر متمایز شدند. سپس جملات پرسشی و تعجبی فیلتر شدند تا انتخاب تنها از بین جملات خبری صورت گیرد. بعد از این مرحله، فیلترکردن براساس طول جملات یعنی پارامتر کمینه و بیشینه، بین ۳ تا ۱۷ کلمه اعمال شد. در نهایت، الگوریتمی با معیارهای مورد نظر پیاده‌سازی شد تا جملاتی را استخراج کند که دربردارندهٔ کلیهٔ دوآوایی‌ها، هجاهای مختلف در جایگاه‌های ابتدای جمله، قبل از نقطه، قبل از کاما، قبل از دونقطه و در سایر جایگاه‌ها و نیز ده هزار کلمهٔ پرکاربرد پیکرهٔ متنی باشد. تمامی انتخاب‌ها از ستون صورت آواییِ پیکره انجام گرفت و تنها در مرحلهٔ جداسازی قطعه‌ها، از متن (صورت نوشتاری) نیز استفاده شد. الگوریتم مذکور به گونه‌ای نمونه جملات را انتخاب می‌کند که هجا یا دوآوایی مورد نظر درون کلمه‌ای انتخاب می‌شود که آن کلمه دارای بسامد رخداد بیشتری است. به عبارتی دیگر، اگر الگوریتم برای انتخاب یک هجا در دو کلمه تلاش می‌کند، اولویت با کلمه‌ای است که دارای بسامد رخداد بیشتری است. انتخاب کلمات با بسامد بالاتر نه تنها به دلیل کاربرد گستردهٔ آنهاست بلکه خواندن آنها برای گوینده نیز ساده‌تر است. سرانجام، ۶۴۱۵ نمونه جمله بدست آمد. این

---

۱. به منظور طراحیِ هدفمند پایگاه‌دادگان، در دسترس داشتن پیکرهٔ متنی مناسب ضروری است. در پژوهش حاضر از یک پیکرهٔ متنیِ ده‌میلیون کلمه‌ای استفاده شده که علاوه بر متن، دارای برچسب‌های اجزای کلام (part of speech (POS))، صورت آوایی و الگوی تکیه نیز می‌باشد. لازم به ذکر است که این پیکرهٔ متنی خود منشعب از یک پیکرهٔ صدمیلیون کلمه‌ای (بی‌جن‌خان، ۱۳۸۶) است که تنها حاوی متن است و پنج گونهٔ زبانی (معیار- رسمی، معیار- غیررسمی، فوق‌معیار- رسمی، فوق‌معیار- غیررسمی و زیرمعیار- غیررسمی) را دربردارد و شامل ۳۵۰۵۸ پروندهٔ متنی است. هر پرونده شامل یک متن کامل یا نمونهٔ تصادفیِ منتخب از یک متن کامل است که بر اساس معیارهای غیرزبانی و همچنین تعلق‌شان به گونه‌های فارسی معاصر انتخاب شده‌اند (همان: ۲۴).

نمونه جملات در مراحل متعددی به صورت دستی اصلاح شدند. برای نمونه، در جملات استخراج‌شده برخی از کلمات فاقد صورت آوایی بودند. در نتیجه صورت آوایی این کلمات به صورت دستی وارد شد. کلمات و عباراتی که به نظر می‌رسید برای خواندن دشوار هستند حذف شدند و از آنجایی که پیکرهٔ متنی متشکل از متونی با ژانرها و سبک‌های مختلف بود، برخی از جملات نامناسب نیز به‌طور کامل حذف شدند. برخی از عبارات تنها به اختصار آمده بودند؛ برای مثال «ص» یا «ر.ک» و ... که همه موارد به صورت کاملشان تبدیل شدند. در نهایت پس از چند مرحله اصلاحات دستی، ۶۴۱۵ نمونه‌جمله استخراج‌شده به ۲۸۲۶ مورد تقلیل یافت. در پایان، نمونه‌جملات اعراب‌گذاری شدند تا خواندن آنها برای گوینده (با توجه به مشکلات ذاتی خط فارسی) تسهیل شود.

انتخاب گوینده

یکی دیگر از مشکلاتی که محققان در ساخت پایگاه‌دادگان با آن روبه‌رو هستند، انتخاب گویندهٔ مناسب است که در این پژوهش، برای به حداقل رساندن تبعات ناشی از آن، ماه‌ها زمان برای گزینش گوینده صرف شد. در ابتدا، از گروهی متشکل از نُه گویندهٔ حرفه‌ایِ صدا دعوت به عمل آمد تا بهترین گزینهٔ ممکن انتخاب شود. تمامی این افراد شاغل در رادیو و در زمان انجام ضبطِ آزمایشی صدا، حداقل سه و حداکثر بیست سال تجربهٔ کاری در این زمینه داشتند. همچنین، دامنهٔ سنی آنها بین ۲۵ تا ۴۵ سال بوده و سبکی که در آن فعالیت داشتند سبک اخباری است. از هر یک از گوینده‌ها خواسته شد تا در یک جلسهٔ ضبطِ آزمایشی حضور پیدا کنند و حدود پنجاه مورد از نمونه جملات، که برحسب تصادف انتخاب شده بود را تولید کنند. این نمونه صداها در اختیار هشت نفر از متخصصانی که در زمینهٔ پردازش گفتار و آواشناسی فعالیت می‌کنند قرار گرفت تا نظر خود را پیرامون مناسب‌ترین صدا، با در نظر گرفتن تمامی معیارهای صدای مناسب برای سیستم‌های بازسازیِ گفتار، اعلام کنند. معیار اصلی برای انتخاب، خوشایندی صدا، قوی بودن صدا، تغییرناپذیری صدا و همچنین قدرت تعامل و همکاری گوینده با تیم تحقیقاتی بوده است. در نهایت، گزینهٔ مناسب از بین این افراد انتخاب شد تا ضبط صدا آغاز شود. در بخش بعدی به فرایند ضبط صدا پرداخته شده است.

ضبط صدا و بررسی فایل‌های صوتی

برخلاف پایگاه‌دادگان‌های مختص سیستم‌های بازشناسی گفتار که گاه در هنگام ضبط، نوفه در محیط جاری است، پایگاه‌دادگان‌های مختص سیستم‌های تبدیل متن به گفتار باید عاری از هرگونه نوفه باشد و باکیفیت‌ترین صدای ممکن ضبط شود (بلک و کُمینک، ۲۰۰۳). برای رسیدن به این هدف یعنی صدای مطلوب، فرایند ضبط صدا که ده جلسه به طول انجامید در اتاقک عایق صدا انجام گرفت و در تمامی جلسات یک آواشناس و یک متخصص فناوری صدا به‌عنوان ناظرِ ضبط حضور داشته‌اند تا در صورت هرگونه خطا در فرایند ضبط، آن را برطرف کنند. فایل‌های صوتی به صورت مونو و با نرخ نمونه‌برداری ۴۴ کیلوهرتز ضبط شده و سبک خواندن جملات توسط گوینده، گفتار اخباری (read-out speech) است. دهان گوینده در

فاصلهٔ بیست سانتیمتری از میکروفون قرار داشت و در این فاصله فیلتری قرار داشت تا صدای ناشی از جریان هوا را به حداقل برساند. به طور میانگین در هر جلسه حدود ۴۰۰ نمونه جملهٔ مناسب بدست آمد و به طور متوسط هر جلسه دو ساعت و نیم طول کشیده است. همچنین، در ابتدای هر جلسهٔ ضبط به جز جلسهٔ اول، تعدادی از جملات ضبط‌شدهٔ جلسات پیشین برای گویندهٔ پخش شد تا سبک و شیوهٔ خواندن را برای گوینده تداعی کند. فاصله و جایگاه میکروفون تا دهان گوینده در طول جلسات مختلف حفظ شد و همچنین به گوینده بازگو شد که حجم صدای خود را ثابت نگه دارد به‌گونه‌ای که در طول ضبط، میزان حجم صدا تغییر نکند. پس از اتمام هر جلسهٔ ضبط تک‌تک فایل‌های صوتی مورد بررسی قرار می‌گرفتند تا در صورت هرگونه عیب و نقصی، جلسهٔ بعد بازضبط آنها انجام گیرد. در نهایت، فایل‌های صوتی مهیا شدند. در کنار فایل‌های صوتی، نمونه جملات و صورت آوایی متناظر با آنها نیز از ابتدا در دسترس بوده است. اما از آنجایی که شیوه‌های خواندن متن توسط گوینده‌ها متفاوت است و همچنین تعدد هم‌نویسه‌ها در زبان فارسی بسیار است، اصلاح دستی نهایی لایهٔ صورت‌های آوایی و گاه لایهٔ متن اجتناب‌ناپذیر است. بنابراین، تک‌تک نمونه‌جملات و صورت آوایی آنها در مراحل گوناگون تصحیح دستی شدند تا سه لایهٔ موج صوتی، متن و صورت آوایی یکدست شوند.

آمار پوشش حالت‌های آوایی و نوایی

میزان پوشش حالت‌های مختلف آوایی و نوایی رابطهٔ مستقیمی با میزان کیفیت خروجی یک سیستم بازسازی گفتار دارد. به بیان دیگر، در شرایط برابر، با افزایش پوشش واحدهای آوایی و نوایی، صدایی با کیفیت مطلوب‌تری را می‌توان انتظار داشت. در بخش حاضر، میزان پوشش این واحدها مورد بررسی قرار می‌گیرد و آمار مربوط به هر یک از آنها در مقایسه با پیکرهٔ متنی مبداء ارائه می‌شود. به منظور آمارگیری از لایهٔ صورت آواییِ پایگاه‌دادگان و پیکرهٔ متنی استفاده شده است. لازم به ذکر است که آمار استخراج‌شده از پیکرهٔ متنی تنها برای جملات با طول بین ۳ تا ۵۰ کلمه اعمال شده است.

اگر تعداد واج‌های زبان فارسی را ۲۹ واج شامل ۲۳ همخوان و ۶ واکه و هجاها را سه نوع cv، cvc، cvcc در نظر بگیریم در این صورت، بدون ممنوعیت هم‌نشینی، هجاهای بالقوه فارسی ۷۶۳۱۴ مورد است. اما از این تعداد، کم‌تر از ۶۰۰۰ مورد آن در فارسی فعلیت دارد (اسلامی و همکاران، ۱۳۸۸: ۶).

همان‌طور که در جدول ۱ نشان داده شده است، در کنار ۲۹ واج معمول فارسی، ۶ واجگونهٔ دیگر در نظرگرفته شده است که ۶ واکه در جایگاه تکیه‌بر (درکنار ۶ واکهٔ دیگر که در جایگاه غیرتکیه‌بر بکار رفته‌اند) را شامل می‌شود. در ضمن، واج سکوت نیز به عنوان یک واج مجزا در نظر گرفته شده است. لازم به ذکر است که برای آوانویسی پیکره و پایگاه‌دادگان از الفبای آوانگار بین‌المللی (IPA) پیروی نشده است بلکه، به خاطر سهولت استفاده از صفحه کلید، برخی از نمادها به صورت قراردادی تعریف شده‌اند. برای مثال، برای واکهٔ پیشینِ افتاده و غیرگرد از «/»، برای همخوان انسدادی چاکنای از «@»، همخوان لثوی- کامی سایشی واکدار از «;»، همخوان لثوی- کامی سایشی بی‌واک از نماد «$»، همخوان لثوی- کامی انسدادی-

سایشی واکدار از «j» و برای همخوان لثوی- کامی انسدادی- سایشی بی‌واک از نماد «c» استفاده شده است. همچنین برای واکه‌های تکیه‌بر سعی شد از حروف بزرگِ متناظر با همتای غیرتکیه‌بر آن استفاده شود. برای معادلِ تکیه‌برِ واکهٔ پیشین افتاده و غیرگرد از نماد «%» استفاده شد. همچنین، سکوت یا مکث (pause) به صورت «|» نمایش داده شده است. جدول ۱ الفبای آوانگار مورد استفاده در این تحقیق را به همراه مثال نشان می‌دهد.

جدول ۱: الفبای آوانگار قراردادیِ مورد استفاده در تحقیق

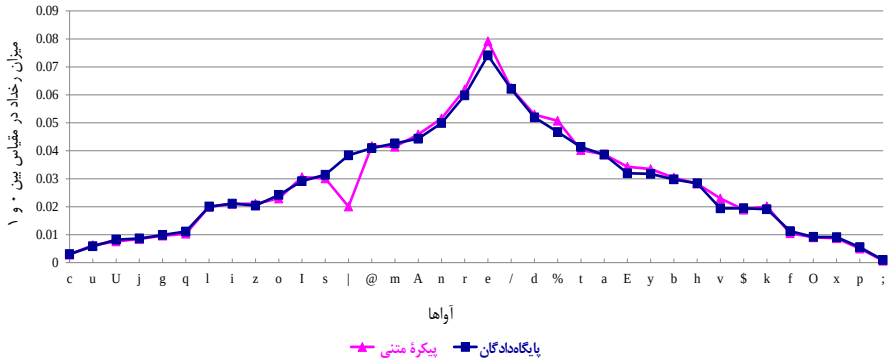
مثال	نماد	مثال	نماد	مثال	نماد	مثال	نماد	مثال	نماد	مثال	نماد	مثال	نماد
است	%	بَرای	/	جام	j	متن	m	لازم	l	باز	b
نیمه	E	شِرکت	e	چنین	c	نور	n	یاری	y	پرداخت	p
ژاپن	O	پُلیس	o	وین	v	راز	r	ژرژ	;	دور	d
سود	U	تاریخ	a	قرینه	q	فردا	f	شش	$	طرف	t
باد	A	اروپا	u	بعد	@	خط	x	زیر	z	گرما	g
سیر	I	ایجاد	i	(سکوت)	|	حریم	h	سال	s	کمک	k

مثال (۱) نمونه جمله‌ای از پایگاه‌دادگان است که به همراه صورت آوایی آن در زیر آمده است:

۱. او سال‌ها در نجف زندگی کرد.

@U salhA d%r n/j%f zendegI k%rd |

در ادامه، در شکل (۱) هیستوگرمِ میزان پوشش واج‌ها در پایگاه‌دادگان در مقایسه با میزان پوشش آنها در کل پیکرهٔ متنی به تصویر درآمده است. محور افقی بیانگر واج‌ها و محور عمودی، میزان رخداد آنها را نشان می‌دهد. همان‌طور که مشاهده می‌شود، میزان نسبت هر یک از واج‌ها در مقایسه با واج‌های کل پیکرهٔ متنی (لایهٔ صورت آوایی) و میزان نسبت همان واج‌ها در پایگاه‌دادگان در مقایسه با واج‌های کل پایگاه‌دادگان تقریباً برابر است. برای مثال، اگر بسامد رخداد واج «O» را در پیکرهٔ متنی بر بسامد رخداد کل واج‌های پیکره تقسیم کنیم به عددی دست پیدا می‌کنیم که بین صفر و یک قرار دارد. اگر همین روند را برای واج «O» در پایگاه‌دادگان طی کنیم نیز به عدد مشخصی بین صفر و یک دست پیدا می‌کنیم که این مقدار برای این واج به ترتیب در پیکرهٔ متنی و پایگاه‌دادگان ۰٫۰۰۹۰۷۶ و ۰٫۰۰۹۱۹۲ است. این مقادیر مشخص برای کلیهٔ واج‌ها بدست آمده است. همان‌طور که در شکل ۱ پیداست تنها تفاوت آشکار در میزان رخدادِ واج‌ها به واج سکوت برمی‌گردد که رخداد آن در پایگاه‌دادگان بسیار بیشتر از پیکرهٔ متنی است.



شکل ۱: هیستوگرم واج‌ها در پایگاه‌دادگان و پیکره

دلیل این تفاوت در این نکته نهفته است که الگوریتمِ انتخاب نمونه جملاتِ حاوی واحدهای نوایی، به گونه‌ای است که بر اساس علائم نقطه‌گذاری مانند نقطه و کاما عمل می‌کند و هر یک از این علائم نقطه‌گذاری، با نماد «|» نمایش داده شده است. بنابراین به دلیل تلاش برای پوشش بیشترین تعداد واحدهای نوایی در این مجموعه از جملات، میزان تعداد سکوت‌ها بیشتر از رخداد سکوت در پیکرهٔ متنی است. در توضیح واکهٔ مرکب «ow» در کلماتی مانند «فردوسی» باید اشاره کرد که این واکه به صورت «w» نمایش داده شده است و در پایگاه‌دادگان نیز لحاظ شده است اما به خاطر پایین بودن رخداد آن در محاسبات آماری از آن صرف‌نظر شده است.

در جدول ۲ در زیر، مجموعهٔ آماری از میزان پوشش کل پایگاه‌دادگان و پیکره آمده است. در این جدول، اطلاعاتی مانند تعداد کل واج‌ها، دوآوایی‌ها، هجاها در جایگاه‌های مختلف، کلمات، جملات و میانگین تعداد کلمات در جمله ارائه شده است. همان‌طور که مشاهده می‌شود علاوه بر اطلاعات عناصر پوشش داده شده در حالت کلی، به میزان پوشش برخی عناصر با (و یا بدون) اعمال واکهٔ تکیه‌بر نیز اشاره شده است.

جدول ۲: آمار کلی میزان پوشش واحدهای مختلف در پایگاه‌دادگان و پیکرهٔ متنی مبداء

واحدهای پوشش داده شده	تعداد (#)					
	کلی		در جایگاه تکیه‌بر		در جایگاه بدون تکیه	
	پایگاه	پیکره	پایگاه	پیکره	پایگاه	پیکره
# کل واج‌ها (قطعه)	۱۹۲۳۳۸	۳۳۷۶۳۲۳۱				
# کل واج‌ها (نوع (type))			۳۶	۳۶	۳۰	۳۰
# دوآوایی‌ها (قطعه)	۱۸۹۵۶۵	۳۳۴۰۳۳۹۶				
# دوآوایی‌ها (نوع)			۱۰۸۶	۱۱۹۳	۸۰۹	۸۵۲

# هجاها (قطعه)	۸۳۴۴۳	۱۴۴۷۴۵۵۷					
# هجاها (نوع)			۳۹۷۵	۶۳۸۱	۲۷۴۰	۳۸۸۲	
# هجاهای ابتدای جمله (نوع)			۹۵۹	۲۳۶۶	۷۹۹	۱۶۸۴	
# هجاهای قبل از دونقطه (قطعه)	۳۸۶	۴۸۸۸۱					
# هجاهای قبل از دونقطه (نوع)			۲۸۴	۱۰۹۵	۲۷۶	۱۰۰۶	
# هجاهای قبل از نقطه (نوع)			۶۰۵	۱۸۴۴	۵۴۰	۱۵۵۶	
# هجاهای قبل از کاما (قطعه)	۳۹۹۲	۲۲۹۳۳۱					
# هجاهای قبل از کاما (نوع)			۱۲۶۳	۲۸۴۶	۱۱۸۹	۲۵۵۸	
# هجاها در سایر جایگاه‌ها (قطعه)	۷۰۶۹۲	۱۳۳۷۲۸۳۶					
# هجاها در سایر جایگاه‌ها (نوع)			۳۳۶۲	۶۳۰۸	۲۳۶۸	۳۸۳۱	
# کل کلمات (قطعه)	۴۰۰۰۳	۶۱۲۷۶۶۵					
# کل کلمات (نوع)					۹۴۵۲	۱۱۰۳۳۲	
# کل جملات	۲۸۲۶	۲۷۷۸۷۴					
میانگین تعداد کلمه در هر جمله	۱۴,۱	۲۲,۰۵					

اما به منظور بررسی تفاوت بین رخداد واحدها در پایگاه در مقایسه با رخداد همتای آن در پیکره از واگرایی کال‌بک لیبلر (Kullback Leibler divergence (KL)) یا همان آنتروپی نسبی (relative entropy) استفاده شده است. کال‌بک لیبلر مقداری نامتقارن از تفاوت بین دو توزیع احتمالی بین Q و P است. به‌طور مشخص، واگرایی Q از P، مقدار اطلاعات از دست رفته است، زمانی که توزیع Q توزیع P را تقریب می‌زند. این نوع واگرایی از طریق فرمول (۱) محاسبه می‌شود. W مجموعه‌ای شامل تمام کلمات ممکن را در برمی‌گیرد. w واحدی اختیاری از مجموعهٔ (W) است. P (w) احتمال w مشاهده شده در پایگاه‌دادگان گفتاری است و Q (w) احتمال w مشاهده شده در پیکرهٔ متنی است. $D_{KL}(P||Q)$ واگراییِ کال‌بک لیبلر Q از P است.

$$D_{KL}(P||Q) = \sum_{w\epsilon W} P(w) ln \frac{P(w)}{Q(w)} \qquad (۱)$$

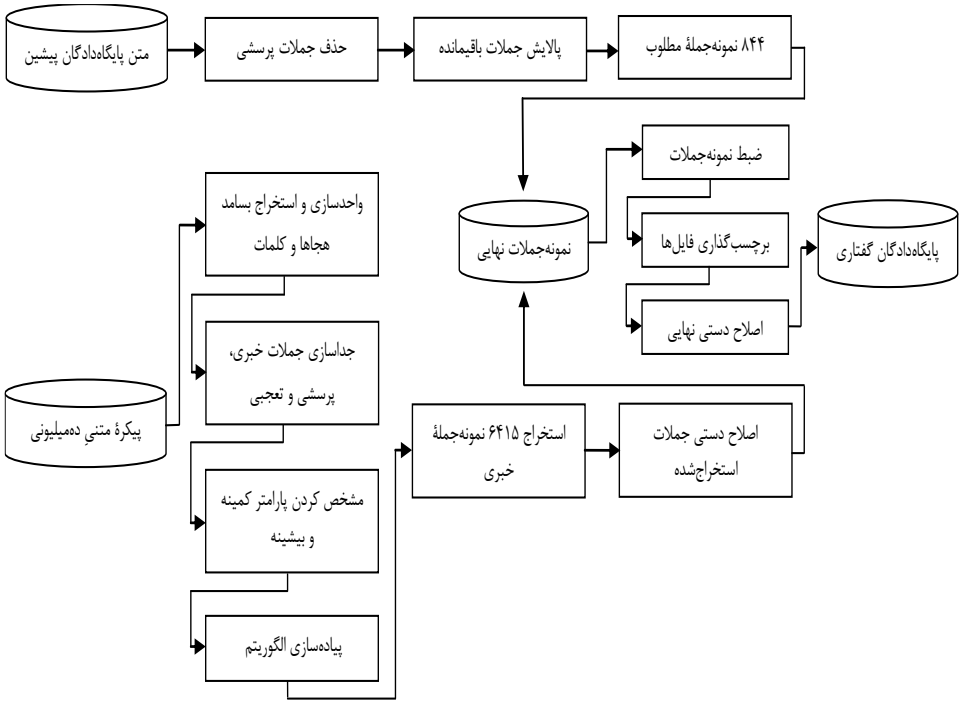
جدول (۳) مقدار KL بدست آمده برای کل واحدها را براساس فرمول (۱) نشان می‌دهد. برپایهٔ این فرمول، KL بدست آمده هر چقدر به سمت صفر نزدیک‌تر باشد حاکی از پوشش مناسب آن واحد در مقایسه با پیکرهٔ متنی مبدا است و بالعکس، به هر میزان که از عدد صفر فاصله بگیرد بیانگر ناکافی بودن میزان آن واحد در پایگاه‌دادگان است.

جدول ۳: میزان KL بدست آمده برای کل واحدها

واحد	مقدار KL	واحد	مقدار KL
دوآوایی، تکیه‌بر	۰/۰۳۴۹	دوآوایی، بدون تکیه	۰/۰۳۱۱
واج، بدون تکیه	۰/۰۰۷۶	واج، تکیه‌بر	۰/۰۰۷۷
هجا، تکیه‌بر	۰/۰۹۸۸	هجا، بدون تکیه	۰/۰۷۵۹
هجای قبل از دونقطه، تکیه‌بر	۲/۰۱۰۶	هجای قبل از دونقطه، بدون تکیه	۱/۹۶۱۰
هجای قبل از نقطه، تکیه‌بر	۰/۴۳۲۱	هجای قبل از نقطه، بدون تکیه	۰/۳۹۱۴
هجای قبل از کاما، تکیه‌بر	۰/۵۹۹۰	هجای قبل از کاما، بدون تکیه	۰/۵۶۱۶
هجا ابتدایی جمله، تکیه‌بر	۰/۶۹۲۸	هجا ابتدایی جمله، بدون تکیه	۰/۵۹۶۵
هجا در سایر جایگاه‌ها، تکیه‌بر	۰/۱۰۲۸	هجا در سایر جایگاه‌ها، بدون تکیه	۰/۰۷۶۸

مقدار KL بدست آمده از پوشش واحدهای مختلف نشان می‌دهد که در بین واحدها، کمترین و بیشترین تفاوت (واگرایی) Q (واحد پوشش داده شده در پایگاه‌دادگان) از P (همتای آن واحد در پیکرهٔ متنی) را به ترتیب، واج‌ها (بدون تکیه) و هجای قبل از دونقطه (تکیه‌بر) داراست. به ترتیب بعد از واج (بدون اعمال تکیه)، بهترین واحد پوشش داده شده عبارتند از دوآوایی (بدون تکیه)، دوآوایی (تکیه‌بر)، هجای (بدون تکیه) و هجای (تکیه‌بر) است. به‌طور کلی هجای قبل از دو نقطه، به طور مناسبی پوشش داده نشده است و می‌تواند حاکی از این نکته باشد که در سیستم بازسازی گفتار نهایی، در مقایسه با دیگر واحدهای نوایی، هجای قبل از دونقطه کیفیت پایین‌تری را داراست.

شکل (۲) طرحوارهٔ مراحل مختلف آماده‌سازی و ضبط پایگاه‌دادگان را نشان می‌دهد.



شکل ۲: طرحوارهٔ مراحل مختلف طراحی و ضبط پایگاه‌دادگان گفتاری مذکور

ساخت صدا و ارزیابی آن

علاوه بر ارائهٔ آماریِ میزان پوشش واحدهای آکوستکی به منظور بررسی میزان دقت و کارآمدی پایگاه‌دادگان‌های گفتاریِ مختص سیستم‌های تبدیل متن به گفتار، بازسازی گفتار با استفاده از آن مجموعه، کارآمدترین معیار برای ارزیابی است. از این رو، با استفاده از دادگانِ پایگاه مذکور به بازسازی صدا با روش فنیِ آماری- پارامتری پرداختیم. در این روش از نوعی میانگین‌گیری از بین واحدهای گفتاری انجام می‌گیرد که در نهایت در زمان بازسازی از آن استفاده می‌شود. در پژوهش حاضر از دو دسته پارامتر یعنی پارامترهای طیفی و پارامترهای تحریک استفاده شده است (زِن و همکاران، ۲۰۰۹). به منظور ساخت این صدا با بکارگیری پارمترهای تحریک، سیگنال تحریک ساخته می‌شود و سپس این سیگنال فیلتر شده است. پس از استخراج پارامترهای مناسب از سیگنال گفتار این پارامترها با یک روش مدل‌سازیِ آماری مدل شدند. سپس، ارزیابی ادراکی صداهای ساخته شده در دستور کار قرار گرفت تا با استفاده از معیارِ میانگینِ امتیازاتِ نظردهی (Mean Opinion Score (MOS)) میزان کیفیت صداهای ساخته شده را ارزیابی کنیم. با استفاده از این معیار، هشت آزمودنی، چهار نفر مرد و چهار نفر زن با میانگین سنی ۲۹ سال از بین عدد ۱ (نازلترین کیفیت) تا ۵ (طبیعی‌ترین کیفیت)، ارزیابی خود را از ۵۰ فایل صوتی بازسازی‌شده که به صورت تصادفی پخش شد

ثبت کردند. به منظور انجام آزمون شنیداری مذکور از ExperimentMFC در برنامهٔ پرت (Praat) استفاده شده است که شنوندگان پس از شنیدن فایل‌های صوتی بازسازی شده، ارزیابی خود را برروی صفحهٔ رایانه ثبت کردند. در نهایت، شنوندگان پس از شنیدن صداهای بازسازی شده کیفیت پاره‌گفتارهای تولیدی را به طور میانگین ۴٫۳ ارزیابی کردند. این میزان ارزیابی، حاکی از کیفیت مطلوب پایگاه‌دادگان طراحی شده است.

نتیجه‌گیری

در مقالهٔ حاضر، به مجموعه مراحلِ طی شده در طراحی و ساخت پایگاه‌دادگان گفتاریِ مختص سیستم‌های تبدیل متن به گفتار فارسی با تمرکز بر پوشش ویژگی‌های نوایی و در نهایت ارزیابی آن از طریق صدای بازسازی‌شده پرداخته شد. پس از آماده‌سازیِ متن و ضبط نمونه جملات با صدای گویندهٔ حرفه‌ای، کلیهٔ صورت‌های نوشتاری و برچسب‌های آوایی اصلاح دستی شدند تا برچسب‌های آوایی با فایل‌های صوتی منطبق شود. در نهایت، صدای بازسازی‌شدهٔ حاصل از این پایگاه‌دادگان و ارزیابی آن نشان‌دهندهٔ کیفیت مطلوبِ مجموعهٔ طراحی‌شده است. در مقایسهٔ این مجموعه با پژوهش‌های پیشین، پوشش طبیعیِ هجاهای مختلف در جایگاه‌های نوایی قابل ذکر است. تاکنون در زبان فارسی پایگاه‌دادگانی با این حجم (پنج ساعت و نیم گفتار) برای سیستم‌های تبدیل متن به گفتار طراحی نشده است. امید است در آینده بتوان به پایگاه‌دادگانی با حجم بیشتر به منظور دستیابی به سیستم‌های با کیفیت مطلوب‌تر نیز دست پیدا کرد. از این مجموعه می‌توان علاوه بر کاربرد اولیه یعنی در سیستم‌های تبدیل متن به گفتار، برای تحلیل‌های آوایی و نوایی فارسی نیز استفاده کرد.

سپاسگزاری

در انجام مراحل مختلف این پروژه افراد زیادی یاری‌رسان بوده‌اند. لازم می‌دانیم مراتب قدرانی خود را به تمامی این عزیزان اعلام داریم: آقایان دکتر ویسی، دکتر بحرانی، کریمی، مشایخی، اسدپور، شریفی و سرکار خانم‌ها بهمنی‌نژاد، نازنینی، چاوشی، معصوم‌زاده، رافعی، شهیدثالث، قادری، نعمتی، تعلیم، اکبری، حبیبی، اسحاق‌تبار، صداقتی و دانشیان. از تک‌تک نامبردگان کمال تشکر را داریم. بی‌شک مسئولیت مطالب مندرج در مقالهٔ فوق برعهدهٔ نگارندگان آن است.

منابع

– اسلامی، محرم؛ شیخ‌زادگان، جواد؛ احمدی‌نیا، زهرا و بهرامی، علی (۱۳۸۸)، مراحل و نحوه تهیه دادگان‌های صوتی هجایی و دایفونی برای سامانه تبدیل متن به گفتار فارسی. دوفصل‌نامه علمی– پژوهشی پردازش علائم و داده‌ها، (۱۲)، ۳–۱۲.

– آیت، سیدسعید (۱۳۸۹)، طراحی و پیاده‌سازی دادگان دایفون زبان فارسی برای کاربرد زبان‌شناسی رایانه‌ای، پژوهش‌های زبان‌شناسی دانشگاه اصفهان، سال دوم، پاییز و زمستان ۱۳۸۹، شماره ۲ (پیاپی ۳)، ۱ـ۱۱.
– بی‌جن‌خان، محمود (۱۳۸۶)، مطالعه و تحقیق جهت تدوین پژوهشنامه عملیاتی دادگان: پیاده سازی استاندارد ایگلز در پیکرهٔ متنی زبان فارسی معاصر، دبیرخانهٔ شورای عالی اطلاع‌رسانی.
– طاهری اردلی، مرتضی و خرم، سهیل (۱۳۹۱)، مدل‌سازی نوای گفتار در سیستم‌های سنتز گفتار فارسی، مجموعه مقالات هشتمین همایش زبانشناسی ایران، به کوشش محمد دبیرمقدم، تهران: دانشگاه علامه طباطبایی، ۴۸۰–۴۹۲.
– همایون‌پور، محمدمهدی (۱۳۹۱)، پژوهشنامهٔ تبدیل متن به گفتار، تهران: شورای عالی اطلاع رسانی، دبیرخانه.
– Abolhasanizadeh, V., Bijankhan, M., & Gussenhoven, C. (2012), The Persian pitch accent and its retention after the focus. Lingua, 122(13), 1380-1394.
– Black, A. (2006), CLUSTERGEN: A Statistical Parametric Synthesizer Using Trajectory Modeling. In: Proc. Interspeech, 1762–1765.
– Black A. W., Zen H. & K. Tokuda (2007), Statistical Parametric Speech Synthesis, ICASSP'2007, pp. IV-1229-IV-1232, Honolulu, Hawai'i, USA.
– Campbell, N. (2005), Developments in corpus-based speech synthesis: Approaching natural conversational speech. IEICE transactions on information and systems, 88(3), 376-383.
– Heusinger, K. (1999), Intonation and information structure. Habilitationsschrift, University of Konstanz.
– Hunt, A., & Black, A. (1996), Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database. In: Proc. ICASSP, 373-376.
– Jurafsky, D., & Martin, J. H. (2007), Speech and language processing. Pearson Education India.
– Khorram, S., Sameti, H., Bahmaninezhad, F., King, S., & Drugman, T. (2014), Context-dependent acoustic modeling based on hidden maximum entropy model for statistical parametric speech synthesis. EURASIP Journal on Audio, Speech, and Music Processing, 2014(1), 12.
– Kominek, J., & Black, A. (2003), CMU ARCTIC databases for speech synthesis. CMU Language Technologies Institute, Tech Report CMU-LTI-03-177.
– Ling, Z.-H., Wang, R.-H. (2006), HMM-based unit selection using frame sized speech segments. In: Proc. Interspeech. 2034-2037.
– Matoušek, J., Tihelka, D., & Romportl, J. (2008), Building of a speech corpus optimized for unit selection TTS synthesis. In Proceedings of LREC 2008, Marrakech, Morocco.
– Moulines, E., Charpentier, F. (1990), Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. Speech Communication, 9, 453-467.
– Nespor, M., & Vogel, I. (2007), Prosodic phonology: with a new foreword (Vol. 28). Walter de Gruyter.
– Sadat-Tehrani, N. (2007), Intonational grammar of Persian, Doctoral dissertation. Manitoba: University of Manitoba.
– Taheri-Ardali, M. & Y. Xu (2012), "Phonetic realization of prosodic focus in Persian". Speech Prosody 2012, Shanghai.
– Taylor, P. (2009), Text-to-speech synthesis. Cambridge, Cambridge University Press.

– Zen, H., Toda, T., Nakamura, M., Tokuda, T., (2007), Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005. IEICE Trans. Inf. Syst, E90-D (1), 325-333.
– Zen H., Tokuda K. and A. W. Black (2009), Statistical Parametric Speech Synthesis, Speech Communication Elsevier, 51(11), 1039-1064.