دانشکده مهندسی برق و رباتیک

گروه الکترونیک

پایان‌نامه کارشناسی ارشد

استخراج دایفون از حالت‌های مختلف گفتاری برای بهبود کارایی سیستم سنتز گفتار

محمدرضا حقیقت

استاد راهنما:

دکتر هادی گرایلو

آبان ماه ۱۳۹۴

دانشکده مهندسی برق و رباتیک
گروه الکترونیک

پایان نامه کارشناسی ارشد

استخراج دایفون از حالت‌های مختلف گفتاری برای بهبود کارایی سیستم سنتز گفتار

محمدرضا حقیقت

استاد راهنما :
دکتر هادی گرایلو

استاد مشاور:
دکتر حسین مروی

آبان ماه ۱۳۹۴

دانشگاه شاهرود

دانشکده برق و رباتیک

گروه الکترونیک

پایان نامه کارشناسی ارشد آقای / خانم .....

تحت عنوان: استخراج دایفون از حالت‌های مختلف گفتاری برای بهبود کارایی سیستم سنتز گفتار

در تاریخ ..... توسط کمیته تخصصی زیر جهت اخذ مدرک کارشناسی ارشد مورد ارزیابی و با درجه ..... مورد پذیرش قرار گرفت.

اساتید راهنما	امضاء	اساتید مشاور	امضاء
نام و نام خانوادگی :		نام و نام خانوادگی :	
نام و نام خانوادگی :		نام و نام خانوادگی :	

اساتید داور	امضاء	نماینده تحصیلات تکمیلی	امضاء
نام و نام خانوادگی :		نام و نام خانوادگی :	
نام و نام خانوادگی :			
نام و نام خانوادگی :			
نام و نام خانوادگی :			

تشکر و قدردانی

از خداوند مهربان تشکر می‌کنم، که با تمام نافرمانی‌هایم ، درهمه‌ی لحظات زندگی یار و یاورم بوده است.

از پدر و مادر عزیزم تشکر می‌کنم ، که همیشه پشتوانه و مشوقم بوده‌اند.

از تمام دوستانم تشکر می‌کنم ، که در تهیه‌ی پایان‌نامه کمکم کرده‌اند.

از استاد راهنما و مشاورم تشکر می‌کنم، که با تمام درگیری‌های خود، بی‌دریغ یار و یاورم بودند.

و

از همسر مهربان و دلسوزم کمال تشکر را دارم، که در تمام لحظات تهیه‌ی پایان‌نامه کنارم بوده و از هیچ تلاش و مساعدتی برای پیشرفتم مضایقه نکرده است.

تعهد نامه

اینجانب ..... دانشجوی دوره کارشناسی ارشد رشته

..... دانشکده ..... دانشگاه صنعتی شاهرود

نویسنده پایان نامه .....

.....تحت راهنمائی.....متعهد می شوم .

- تحقیقات در این پایان نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است .
- در استفاده از نتایج پژوهشهای محققان دیگر به مرجع مورد استفاده استناد شده است .
- مطالب مندرج در پایان نامه تاکنون توسط خود یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است .
- کلیه حقوق معنوی این اثر متعلق به دانشگاه صنعتی شاهرود می باشد و مقالات مستخرج با نام « دانشگاه صنعتی شاهرود » و یا « Shahrood University of Technology » به چاپ خواهد رسید .
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان نامه تأثیرگذار بوده اند در مقالات مستخرج از پایان نامه رعایت می گردد.
- در کلیه مراحل انجام این پایان نامه ، در مواردی که از موجود زنده (یا بافتهای آن‌ها) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است .
- در کلیه مراحل انجام این پایان نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده است اصل رازداری ضوابط و اصول اخلاق انسانی رعایت شده است .

تاریخ

امضای دانشجو

مالکیت نتایج و حق نشر

- کلیه حقوق معنوی این اثر و محصولات آن (مقالات مستخرج ، کتاب ، برنامه های رایانه ای ، نرم افزار ها و تجهیزات ساخته شده است) متعلق به دانشگاه صنعتی شاهرود می باشد . این مطلب باید به نحو مقتضی در تولیدات علمی مربوطه ذکر شود .
- استفاده از اطلاعات و نتایج موجود در پایان نامه بدون ذکر مرجع مجاز نمی باشد.

چکیده

در این پایان‌نامه، دادگان جدیدی با استفاده از دایفون‌ها برای سیستم سنتز گفتار پیشنهاد داده‌ایم. دایفون‌ها چند مزیت بزرگ نسبت به سایر واحدهای گفتار دارند. یکی این است که تعداد آن‌ها کم است و دیگر این‌که دایفون‌ها گذر از یک دایفون به دایفون بعدی را نسبت به واحدهای دیگر گفتار، بهتر نشان می‌دهند.

در این تحقیق می‌خواهیم حالت‌های مختلف جملات را در خروجی سیستم سنتز گفتار تولید کنیم. در بیشتر روش‌های استفاده شده در زبان فارسی، برای طبیعی‌سازی گفتار مصنوعی از مدل‌های آماری و ریاضی مانند مدل مخفی مارکوف، درخت‌های تصمیم‌گیری، شبکه‌های عصبی و ... استفاده شده است که به‌کارگیری آن‌ها بسیار پیچیده می‌باشد. در این پایان‌نامه نشان داده می‌شود که می‌توان با در دست داشتن حالت‌های مختلف دایفونی گفتار سنتزشده‌ی خروجی را با حالت‌های مختلف گفتاری، تولید، و آن را به گفتار طبیعی انسان نزدیک‌تر کرد. دادگان دایفونی پیشنهاد شده، شامل هفت حالت مختلف دایفونی است که هر کدام دارای ویژگی منحصر به فردی می‌باشد.

در بخش شبیه‌سازی، ابتدا با استفاده از دایفون‌های استخراجی از جملات ضبط شده و چسباندن دایفون‌های مورد نیاز در کنار هم، جملات هدف را تولید کردیم. جملات به دست آمده از این روش را برای شنوندگان پخش نمودیم و ایشان طبق معیار MOS نمره‌ای برای پارامترهایی مانند قابلیت فهم و طبیعی بودن گفتار لحاظ کردند. مقایسه نتایج بدست آمده با نتایج مقالات دیگر که طبق معیار MOS روش خود را ارزیابی کرده‌اند نشان میدهد که روش پیشنهادی از نظر قابلیت فهم امتیاز بیشتری نسبت به روش استفاده از درخت تصمیم‌گیری و روش استفاده از ترایفون کسب نموده است. همچنین راه‌کارهایی برای ارتقای کیفیت روان بودن و طبیعی بودن گفتار خروجی، وجود دارد که در پایان‌نامه به شرح آن‌ها پرداخته‌ایم.

واژگان کلیدی: سیستم تبدیل متن به گفتار، سیستم سنتز گفتار، دایفون، طبیعی‌سازی گفتار مصنوعی، معیار MOS

فهرست مطالب

فهرست شکل‌ها

فهرست جداول

فصل اول:

مقدمه

تولید سیستم‌های متن به گفتار و ارتقای آن برای زبان فارسی، حدود دو دهه است که دغدغه‌ی بسیاری از محققین شده است. این سیستم دارای بخش‌های مختلفی است که هر کدام از آن‌ها تخصصی است و از نظر تئوری با هم متفاوت می‌باشند. در این پایان‌نامه می‌خواهیم تحقیق خود را در قسمتی از سیستم سنتز گفتار انجام دهیم.

با توجه به طبیعی نبودن گفتار تولید شده از سیستم سنتز، به دنبال روشی برای نزدیک کردن آن به گفتار طبیعی انسان هستیم. با تحقیق در کتب زبان‌شناسی و پردازش گفتار و مرور روش‌های موجود، روش جدیدی ارائه دادیم و آن را شبیه‌سازی کردیم و با بقیه‌ی روش‌ها از نظر کیفی و کمی، مقایسه انجام دادیم.

در روش‌های موجود دو نکته وجود دارد: یکی نوع واحدهای گفتار و دیگری روش طبیعی‌سازی خروجی. در بیشتر روش‌ها از هجا به عنوان واحد گفتار استفاده شده است ولی هجاها نمی‌توانند به خوبی گذر از یک هجا به هجای دیگر را نشان دهند و این امر باعث می‌شود که بعد از سنتز آن‌ها، گفتار خروجی حالت کوبنده‌ای داشته باشد و تغییرات ناگهانی شدت صدا را در خروجی شاهد خواهیم بود. برای رفع این مشکل از واحدهای کوچکتری به نام دایفون استفاده کردیم که تا حد خیلی زیادی این مشکل را برطرف کرده است. برای طبیعی‌سازی گفتار نیز بیشتر روش‌ها و مقالات، مدل‌های مختلف ریاضی و آماری را پیشنهاد داده‌اند، که در عین حال که استفاده از آن‌ها بسیار سخت و پیچیده است، خروجی چندان خوبی ندارند. برای حل این مشکل نیز از انواع دایفون در حالت‌های مختلف گفتاری استفاده کردیم. هفت حالت مختلف برای دایفون پیشنهاد شده است، که بتوان جملات را هنگام سنتز با تکیه و آهنگ درست ادا کرد. این هفت حالت عبارت‌اند از:

- دایفون‌های بدون تکیه
- دایفون‌های تکیه‌دار
- دایفون‌های هجای ابتدایی فعل پرسشی با واژه‌ی "آیا"

- دایفون‌های هجای ابتدایی فعل پرسشی ساده یا فعل خبری
- دایفون‌های هجای ابتدایی فعل تعجبی
- دایفون‌های هجای انتهایی فعل پرسشی با واژه‌ی "آیا"
- دایفون‌های میان‌واژه‌ای

در فصل دوم درباره‌ی کارهای انجام شده در زمینه‌ی سنتز گفتار و روش‌های ارزیابی آن مطالبی آورده‌ایم. همچنین روش‌های مقالات مختلف برای طبیعی‌سازی گفتار خروجی حاصله از سیستم سنتز را در انتهای این فصل بررسی کرده‌ایم.

در فصل سوم مبانی تئوری که برای انجام این پایان‌نامه نیاز بوده است مورد بررسی قرار می‌گیرد. این فصل به دو موضوع مهم زبان‌شناسی و انواع دادگان زبان فارسی اشاره دارد. در انتهای آن فصل نیز روش‌های موجود برای تقطیع و استخراج واحدهای گفتار مورد بررسی قرار گرفته‌اند.

در فصل چهارم توضیحات مفصلی درباره‌ی روش پیشنهادی نگارنده‌ی پایان‌نامه آمده است و استفاده از این روش برای تولید دادگان بزرگ دایفونی زبان فارسی، توصیه شده است. برای تولید یک پایگاه داده‌ی بزرگ دایفونی مسائل مهمی وجود دارد که با رعایت آن‌ها می‌توان کیفیت گفتار خروجی سیستم سنتز گفتار را ارتقا داد. در این فصل به تعدادی از این مسائل اشاره شده است و برای رفع هر کدام از آن‌ها راه‌حل‌هایی را پیشنهاد داده‌ایم.

در فصل پنجم نتایج آزمایشات انجام شده با استفاده از روش پیشنهادی را به دست آورده‌ایم. این نتایج را با نتایج یکی از بهترین روش‌های موجود که توسط نگارنده با شرایط یکسان شبیه‌سازی شده است، مقایسه کرده‌ایم. مقایسه‌ی آن‌ها به این صورت است که گفتار حاصله از سنتز را برای شنوندگان پخش می‌کنیم و آن‌ها باید از ۱ تا ۵ نمره‌ای لحاظ کنند. این نمرات یا عدد کامل هستند یا با ۰/۵ رقم اعشار و بهترین نمره ۵ است. شنوندگان به هر جمله باید از سه لحاظ نمره بدهند: "قابل فهم بودن"، "صحت نوای گفتار" و "روانی کلام". این روش ارزیابی را MOS[1] می‌نامند. در نهایت روش

[1] *Mean Opinion Score*

پیشنهادی را با سه روش موجود دیگر که امتیاز آن‌ها با معیار MOS در مقاله‌ی [۱] آمده است، مقایسه می‌کنیم. روش پیشنهادی از نظر "قابل فهم بودن" و "صحت نوای گفتار"، خروجی بسیار خوبی دارد؛ ولی به دلیل عدم هنجارسازی واحدهای دایفونی و مناسب نبودن دایفون‌های استخراج شده توسط نگارنده و بسیاری از شرایط دیگر که می‌بایست رعایت می‌شد و به دلایلی امکان انجام آن-ها ممکن نبود، سنتز خروجی، "روانی کلام" مطلوبی نداشت و امتیاز خیلی خوبی نسبت به دو پارامتر دیگر نگرفت.

در فصل آخر که فصل ششم می‌باشد، نتیجه‌گیری حاصل از این شش فصل به صورت بسیار مختصر صورت گرفته است. هم‌چنین پیشنهادات نگارنده در تیتر بعدی فصل ششم آمده است و استفاده از این روش را در تهیه‌ی دادگان دایفونی زبان فارسی مناسب می‌داند. هم‌چنین پیشنهادات دیگری نیز برای ارتقای سطح سیستم‌های متن به گفتار داده شده است.

فصل دوم:

روش‌های سنتز گفتار و ارزیابی و طبیعی‌سازی آن

۲-۱. مقدمه

در این فصل، راجع به زبان طبیعی و پردازش آن، توضیح مختصری داده‌ایم. این علم بزرگ، شامل زیرمجموعه‌های زیادی از جمله تبدیل متن به گفتار می‌باشد. درباره‌ی سیستم متن به گفتار نیز، توضیح کوتاهی دادیم و به زیرمجموعه‌ی آن که سنتز گفتار است، اشاره کردیم. سنتز گفتار، یکی از روش‌های تولید گفتار مصنوعی است که خود، شامل زیرمجموعه‌های فراوانی است. در این فصل، درباره‌ی واحدهای مورد استفاده برای تهیه‌ی دادگان سیستم متن به گفتار، توضیح داده شده است. یکی از پارامترهای مهمی که در سنتز گفتار باید بدان توجه شود، طبیعی بودن گفتار خروجی می‌باشد. از این رو، مبحث نوای گفتار را شرح و بسط دادیم که با رعایت پارامترهای نوای گفتار، خروجی سنتز (به گوش شنونده) طبیعی و خوشایند خواهد بود. سپس درباره‌ی شیوه‌های سنجش عملکرد سیستم سنتز نیز، توضیح مختصری داده شده است.

در آخر نیز به بررسی روش‌های استفاده شده در تحقیق‌های موجود برای طبیعی‌سازی گفتار سنتز-شده‌ی خروجی می‌پردازیم. با توجه به روش‌های موجود، در فصل روش پیشنهادی برای طبیعی‌سازی سنتز خروجی از روش واحدهای حساس به بافت‌های نوایی (البته با تقسیم‌بندی نوایی نگارنده)، استفاده شده است

۲-۲. زبان طبیعی و پردازش آن

پردازش زبان طبیعی[1] عبارتست از: استفاده از رایانه برای پردازش زبان گفتاری و نوشتاری.

پردازش زبان طبیعی از دو قسمت اصلی تشکیل شده است:

۱- پردازش

۲- زبان طبیعی

[1] *Natural Language Processing*

بنابراین برای فهم بیشتر این دانش، به دو قسمت از علوم نیاز است: علوم کامپیوتر در بحث پردازش، و علوم زبان‌شناسی در بحث زبان طبیعی. یعنی علاوه بر فعالیت در حوزه‌ی پردازشی، باید فعالیت‌هایی در زمینه‌ی شناخت زبان طبیعی (که همان زبانی است که با آن صحبت می‌کنیم و می‌نویسیم) نیز، صورت گیرد.

زبان طبیعی مشخصه‌های مختلفی دارد، مثلا ۴ ویژگی از آن در ادامه مطرح شده است:

- زبان طبیعی زبانی است که در تعاملات اجتماعی روزمره، با استفاده از آن می‌نویسیم و صحبت می‌کنیم.
- زبان‌های طبیعی مختلف و زیادی وجود دارند. فارسی، انگلیسی، ژاپنی، چینی، ...
- ممکن است که فرم گفتاری و نوشتاری زبان‌ها متفاوت باشند و یا از هم مستقل باشند.
- برخی از ویژگی‌های زبان را می‌توان به صورت "مدل" ارائه کرد.

پردازش زبان طبیعی، می‌تواند در سطوح مختلف از زبان صورت گیرد. برای زبان طبیعی می‌توان مقوله‌های زیر را در نظر گرفت[۲]:

۱) آواشناسی و صداشناسی[1]: که به تشخیص آواها و صداها و بازشناسی گفتار می‌پردازد.

۲) ریخت‌شناسی[2]: که به ساختار کلمات و ریشه‌یابی واژگان می‌پردازد.

۳) نحو[3]: که به ارتباط کلمات به هم‌دیگر، و مباحث دستوری آن‌ها در گروه‌ها و جملات می‌پردازد.

۴) معناشناسی[4]: که به ارتباطات معنایی کلمات می‌پردازد.

۵) عمل‌گرایی[5]: که به کاربردهای زبان برای رساندن یک مطلب به مخاطب یا مخاطبان، در حالت عملی و یا در نوشتار و گفتار طبیعی می‌پردازد.

---

[1] *Phonetics and Phonology*

[2] *Morphology*
[3] *Syntax*
[4] *Semantics*
[5] *Pragmatics*

۶) مباحثه[1]: که به ارتباطات کلی یک زبان، فرای یک یا چند جمله‌ی خاص می‌پردازد.

در علوم کامپیوتر نیز، می‌توان به نیازمندی‌های زیر اشاره کرد:

- پردازش متن
- پردازش گفتار

که هر کدام از آن‌ها، به دو قسمت تقسیم می‌شوند :

- شناسایی
- تولید

یعنی رایانه باید بتواند یک متن/گفتار را شناسایی کند و متناسب با آن یک متن/گفتار تولید کند. با این اوصاف، حالت‌های زیر برای یک سیستم پردازش زبان طبیعی، از حیث نوع ورودی و خروجی متصور است:

- ورودی متن – خروجی متن
- ورودی گفتار – خروجی متن
- ورودی متن – خروجی گفتار
- ورودی گفتار – خروجی گفتار

در نهایت می‌توان اینگونه جمع‌بندی کرد که برای دستیابی به یک سیستم پردازش زبان طبیعی قدرتمند، ابتدا باید زبان موردنظر را به‌خوبی شناخت (زبان‌شناسی)، سپس به کمک راه‌کارهای پردازشی (علوم کامپیوتر و محاسباتی) به شناسایی و تولید متن و گفتار پرداخت.

۲-۳. پیش‌پردازش زبان طبیعی

برخی مسائل اساسی در پیش‌پردازش زبان طبیعی را می‌توان به صورت زیر برشمرد[۳] :

- قطعه‌بندی گفتار[1]

[1] *Discourse*

- قطعه‌بندی متن[۲]
- برچسب‌گذاری نقش کلمه[۳]
- ابهام‌زدائی از نقش‌کلمه[۴]
- ابهام‌زدائی نحوی[۵]
- هنجارسازی[۶]
- تشخیص اعمال گفتاری[۷]

مسائل فوق، تقریباً در تمامی کاربردهای پردازش زبان طبیعی در حوزه‌ی خط و زبان مطرح هستند و به عنوان مسائل زیرساختی خط و زبان، باید مورد بررسی قرار گیرند.

۲-۴. بیان ساده‌ای از یک سیستم تبدیل متن به گفتار[۸]

سیستم متن به گفتار را به گونه‌ای در نظر می‌گیریم که ورودی متن را به صورت کاراکترهای اسکی با طول دلخواه بگیرد و با پردازش‌های لازم، خروجی گفتار طبیعی داشته باشد. برای اعمال کنترل بیشتر بر روی این سیستم، ابتدا این متن، به جملات تشکیل‌دهنده‌اش تقسیم می‌شود، که برای این قسمت، یک بلوک تشخیص جمله، نیاز است. سپس شناسایی کلمات، اعداد، تاریخ و دیگر علائم صورت می‌گیرد. علائم و نشانه‌ها با یک بلوک دیگر، ترجمه می‌شوند و به صورت کاراکترهای متنی درمی‌آیند. مثلا "x+y" به صورت "ایکس به علاوه‌ی وای" ترجمه می‌شود. همچنین، رفع ابهام زبانی نیز، باید انجام شود.

---

[1] *Speech Segmentation*
[2] *Text Segmentation*

[3] *Part-of-Speech Tagging*
[4] *Word Sense Disambiguation*
[5] *Syntactic Disambiguation*
[6] *Normalization*
[7] *Speech acts*

[۲۰] این پاراگراف از [۴]و[۵] گرفته شده است.

سپس با استفاده از روش مورد نظر، متن را به صدا تبدیل می‌کنیم. و سپس خروجی را هنجارسازی می‌کنیم تا خروجی طبیعی و دلخواه‌مان به دست آید.

سیستم تبدیل متن به گفتار، کاربردهای زیادی دارد. از جمله‌ی آن‌ها، می‌توان به موارد زیر اشاره کرد:

- در سیستم‌های گویا.
- در تلفن‌بانک‌ها.
- در فرودگاه‌ها.
- مناسب برای افراد نابینا به جای استفاده از خط بریل، یا استفاده از آن در رایانه.
- خواندن متون مختلف برای کاربران رایانه‌ای و اینترنتی.

مسائل اصلی که در یک سیستم متن به گفتار وجود دارد، به ترتیب زیر است:

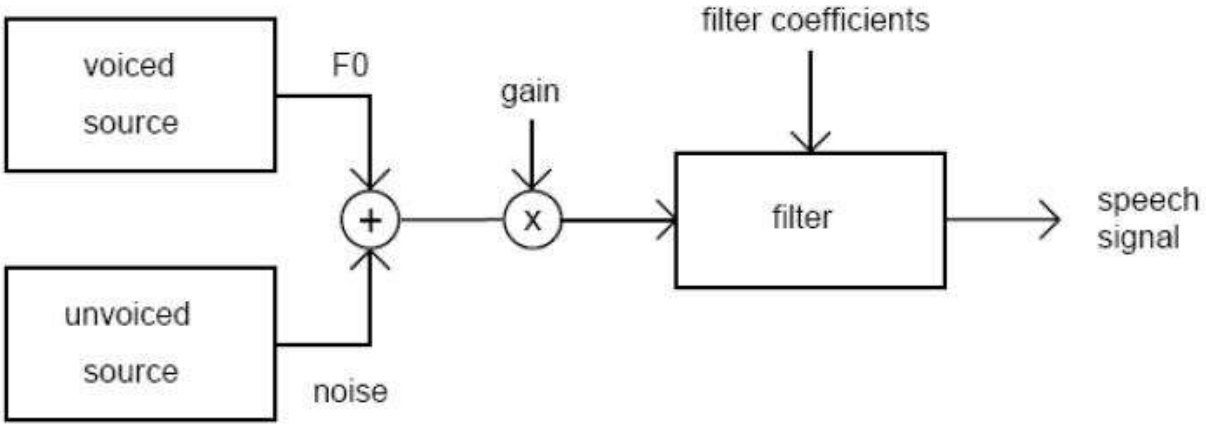
- دسته‌بندی علائم در متن: در تمام زبان‌ها، علائمی وجود دارند که فرم نوشتاری و گفتاری آن‌ها متفاوت است. علائمی مانند: ساعت، تاریخ، علائم ریاضی، اختصارات و ... .
- مشکلات زبان طبیعی: مشکلات و چالش‌های پردازش متن، شامل موارد خیلی زیادی است؛ مثلا ابهام در نقش کلمات یا چندمعنایی‌بودن برخی کلمات یا ... که باید حل شوند.
- تولید صدای طبیعی: گفتار خروجی باید طوری باشد که شنونده نتواند تشخیص دهد که صدای تولیدی، ماشینی است.
- تولید گفتار قابل فهم.
- نوای گفتار.

۲-۵. روش‌های تولید شکل موج گفتار[۱]

گفتار می‌تواند با روش‌های مختلفی تولید شود. هرکدام از آنها، دارای معایب و محاسنی هستند. در ذیل، انواع این روش‌ها به اختصار توضیح داده شده است:

[۱] این قسمت از [۶]و [۷،۸] استفاده شده است.

- تولید گفتار مفصلی: تولید گفتار مفصلی، سعی در مدل‌سازی اعضای صوتی بدن به طور کامل دارد. بنابراین به طور بالقوه بهترین روش برای تبدیل متن به گفتار با کیفیت بالا می‌باشد. اما از آن‌جا که پیاده‌سازی آن مشکل است، نسبت به روش‌های دیگر، کمتر مورد توجه قرار گرفته است.
- تولید گفتار فرمنت: این روش مبتنی بر مدل فیلتر کردن منبع گفتار است، که در دهه‌های گذشته مورد استفاده‌ی بسیار زیادی قرار گرفته است. تولید گفتار فرمنت دارای ساختارهای موازی، سری و ترکیبی است. این روش قابلیت تولید تعداد نامحدودی از صداها را دارد و از این حیث از روش‌های اتصالی، انعطاف‌پذیرتر است. حداقل فرمنت موردنیاز، سه عدد می‌باشد و برای کیفیت بیشتر، از تعداد بیشتری استفاده می‌شود. معمولا هر فرمنت با یک سیستم ارتعاشی دو قطبی، مدل می‌شود؛ که فرکانس فرمنت و پهنای باند آن را مشخص می‌کند. در شکل ۲-۱ کلیت این سیستم نشان داده شده است.



شکل ۲-۱: کلیت روش فرمنت در تبدیل متن به گفتار

ساختار سری برای حروف صدادار بهتر عمل می‌کند و ساختار موازی برای حروفی که وابسته به بینی هستند و صامت‌ها عملکرد خوبی دارند. ساختارهای سری-موازی، در سیستم امروزی کارایی بهتری نسبت به قبلی‌ها دارد.

- تولید گفتار اتصالی یا سنتز گفتار: اتصال تکه‌های صوتی از قبل ضبط شده‌ی طبیعی[1] و تولید گفتار طبیعی، روش کارآمدتر و طبیعی‌تری است. تولیدکننده‌های گفتار اتصالی، حداقل به یک گوینده با یک صدا نیاز دارند و نسبت به سایر روش‌ها حافظه‌ی بیشتری اشغال می‌کنند. واحدهای گفتاری می‌توانند کلمات، هجاها، نیم هجاها و واج‌ها باشند که در مطالب گذشته، در مورد این اجزا توضیح داده شده است.

استفاده از این روش، نارسایی‌هایی در گفتار خروجی خواهد داشت که باید برطرف شود. نوای گفتار و پیوستگی آن، باید تصحیح شود. از این به بعد، بیشتر در مورد مزایا و معایب این روش -که پرکاربردترین روش است- بحث می‌شود.

در [۹،۱۰] از واج و در [۱۱،۱۲] از بخش‌های کلمه، برای بازسازی استفاده شده است.

در روش‌های فوق "سرعت" و "میزان اشغال شدن حافظه" با هم مصالحه دارند.

۲-۶. واحدهای پردازش دادگان

یکی از قسمت‌های مربوط به پردازش زبان طبیعی، در دست داشتن داده‌های زبانی مربوط به یک زبان است. اگر واحد بزرگی نظیر جمله را کنار بگذاریم، قطعات استفاده شده در تهیه‌ی دادگان عبارت‌اند از: واژه، واج، هجا، دایفون و برخی واحدهای دیگر [۴] که در ادامه بدان‌ها پرداخته شده است:

- واژه: بعد از جمله و عبارت، واژه، بزرگترین واحدی است که در تهیه‌ی دادگان استفاده می‌شود، ولی باید توجه داشت که تعداد واژه‌های مورد استفاده در یک زبان بسیار زیاد است؛ مثلاً در زبان انگلیسی حداقل ۴۰۰۰۰ واژه در گفتار روزمره مورد نیاز است. از سویی دیگر، تعداد واژه‌های هر زبان نیز به طور پیوسته در حال تغییر است، و حافظه‌ی فوق العاده زیاد و به‌روزرسانی مکرر دادگان را می‌طلبد. از منظری دیگر، بین واژه‌های ادا شده در یک جمله، نوعی ارتباط و به عبارتی تأثیر متقابل آواها وجود دارد که در صورت استفاده از دادگان

[1] تکه‌های ضبط شده صدای انسان است و نه ماشین. و در هنگام سنتز، همین صداهاست که باهم ترکیب می‌شوند.

واژگان، تأثیر آواها در انتهای واژه و ابتدای واژه‌ی بعد، لحاظ نمی‌شود. مشکلات موجود در واحدی نظیر واژه، محققان را بر آن داشت تا واحدهای آوایی کوچکتری را انتخاب کنند که ضمن نیاز به حافظه‌ی کمتر، کیفیت مناسب‌تری را نیز، امکان‌پذیر سازند.

- واج: واج، عبارت است از واحد اساسی، مجرد، و انتزاعی هر زبان، که برای انتقال معانی به کار می‌رود. واج‌ها، کوچکترین واحدهای آوایی‌اند که تعویض آن‌ها موجب تغییر معنایی واژه می‌گردد. در زبان فارسی، ۲۹ واج و زبان انگلیسی، ۴۰ تا ۵۰ واج وجود دارد ودر مقایسه با واژه‌های زبان، حافظه‌ی کمی برای ذخیره‌ی آن‌ها مورد نیاز است. واج، کوچکترین عنصر ممکن در دادگان است و درهر زبانی، واحد آوایی مشخص، و با تعداد کاملاً مشخص و محدود محسوب می‌شود. گرچه واج‌های یک زبان کم‌اند، ولی به دلیل اثر هم‌آوایی میان واج‌ها، معمولاً تعیین مرز دقیق واج‌ها میسر نیست. از سوی دیگر، در کاربردی نظیر گفتارساز (تولید گفتار مصنوعی توسط ماشین)، قرارگرفتن واجی متفاوت با واجی که هنگام ضبط دادگان استفاده می‌شود، بعد از واج فعلی، اثر متقابل بین واج‌ها را آن‌گونه که در زبان طبیعی وجود دارد، مدل می‌کند. بنابراین، در عمل، در کاربردی نظیر گفتارسازها، واج‌ها به ندرت در تهیه‌ی دادگان استفاده می‌شوند.
- هجا: هجا، از آواهایی تشکیل می‌شود که ساخت و ترکیب آن، بسته به نوع زبان متفاوت است. هجا، رشته‌ی آوایی پیوسته‌ای است؛ یعنی اجزای سازنده‌ی هجا، طی فرایند تولیدی بدون مکث، ادا می‌شوند. برای مثال، هجا در زبان فارسی از یک واکه (مصوت) و یک تا سه همخوان (صامت) تشکیل می‌شود. هجای آغازین حتماً همخوان است و نمی‌تواند واکه باشد. در آغاز هجا نیز، دو همخوان پشت سر هم نمی‌توانند قرار بگیرند؛ در حالی که در زبانی مانند انگلیسی، در کلمه‌ای مانند *street* ابتدا با واکه شروع می‌شود. بنابراین، در زبان فارسی سه نوع هجا وجود دارد که با قرار دادن C به جای همخوان و V به جای واکه، به سه صورت CV

و *CVC* و *CVCC* درمی‌آید. مثال‌های این سه نوع (به ترتیب) عبارت‌است از: تو، من، گوشت. برای تهیه‌ی دادگان هجایی در انگلیسی، تقریبا به ۱۰۰۰۰ هجا نیاز است.

- دایفون[1]: همان‌گونه که پیش‌تر بیان شد، واحد پایه در تهیه‌ی دادگان، باید به گونه‌ای باشد که اولاً، حجم حافظه‌ی معقولی را اشغال کند، به عبارتی، تعداد عناصر دادگان مطلوب باشد؛ و دوم اینکه، بتوان تأثیر متقابل میان آواها را با آن دادگان در نظر گرفت. از جمله واحدهایی که با هدف تأمین این شرایط تعریف و استفاده می‌شوند، واحد دایفون است. این واحد، در واقع برای در نظر گرفتن انتقال از یک واج به واج بعدی ابداع شده است. برای دایفون، تعاریف مختلفی بیان شده است. مثلا: دایفون، عبارت است از نیمه‌ی پایدار یک آوا تا نیمه‌ی پایدار آوای بعدی. برای تهیه‌ی دادگان دایفون در فارسی، تقریبا ۹۰۰ داده و در انگلیسی ۱۰۰۰ داده نیاز است.
- واحدهای‌دیگر: واحدهای دیگری نظیر سه واجی[2] که می‌تواند شامل نیمه‌ی دوم یک واج، یک واج کامل، و نیمه‌ی اول واج بعد شود نیز، وجود دارند که بسته به کاربرد، ممکن است استفاده شوند. با این وجود، دادگان دایفون از مهمترین دادگان‌هایی است که به دلیل در نظر گرفتن مرز میان واج‌ها، می‌تواند در گفتارسازهای پیوندی و سایر مطالعات مرتبط زبان‌شناسی استفاده شود.
- رابطه‌ی بین هجا و دایفون: دایفون، جزئی از یک هجا است. اگر صامت‌ها را (که در زبان فارسی ۳۲ عدد هستند) با C، و مصوت‌ها را (که در زبان فارسی ۶ عدد می‌باشند) با V نمایش دهیم، ساختارهای ممکن در زبان فارسی عبارتند از: CV,CVC,CVCC. اولین تفاوتی که در ساختار دایفون و هجا از نظر الگوی صامت و مصوت وجود دارد، استفاده از ساکن (با نمایش -) در ساختار دایفون‌هاست. با توجه به تعریفی که برای دایفون ارائه شد، ساختار آن به این

[1] *Diphone*
[2] *Triphone*

صورت بیان شده: -C,CV,VC,C-,CC,V-. پس رابطه‌ی بین هجا و دایفون به صورت زیـر، قابـل بیان می‌باشد:

CV: -C/CV/V-

CVC: -C/CV/VC/C-

CVCC: -C/CV/VC/CC/C-

جدول ۲-۱: مقایسه‌ی واحدهای گفتار

واحد گفتاری	تعداد	متوسط دیرش زمانی هر واحد بر حسب ms	حافظه‌ی مورد نیاز	پیچیدگی محاسباتی روش بازسازی
کلمه	۱۰۰هزار تا ۱،۵ میلیون	۳۵۰	زیاد	پایین
واژک	۱۲۰۰۰	۲۵۰	متوسط	پایین
هجا	۱۱۰۰۰	۲۰۰	متوسط	پایین
نیمه‌هجا	۲۰۰۰	۱۰۰	کم	پایین
واج	۴۲	۸۰	کم	پایین
دوآوایی	۱۵۰۰	۸۰	کم	پایین
سه‌آوایی	۳۰۰۰۰	۸۰	متوسط	پایین

۲-۷. نوای گفتار

چهار عنصر نوایی گفتار که معمولا در سطوح مختلف اعم از هجا، واژه و یا جمله اثر خود را نشان مـی-دهند عبارتند از: زیری و بمی یا منحنی گـام[1]، دیـرش[2] (دوره)، انـرژی یـا شـدت[3]، و درنـگ[4]. اعمـال اطلاعات نوای گفتار به سیستم سنتزکننده، نقش بـه سـزایی در تولیـد گفتـار طبیعـی، در زبـان‌هـای

[1] *Pitch*
[2] *Duration*
[3] *intensity*
[4] *Pause*

مختلف دارد. در تحقیقات جدید، پارامتر دیگری نیز مورد توجه قرار گرفته است که فاز[1] نام دارد. در ادامه، به توضیح مختصری با توجه به آن‌چه در [۶،۱۰] آمده است، می‌پردازیم.

- منحنی گام: از بخش‌های اصلی نوای گفتار، منحنی گام می‌باشد که تغییرات آن بیشترین تاثیر را در تولید نوای طبیعی دارد. تولید منحنی گام مناسب، شامل دو بخش می‌باشد؛ ابتدا باید مکان‌های زیر و بمی تشخیص داده شوند (معمولا با استفاده از اطلاعات زبان‌شناسی صورت می‌گیرد) و وابسته به ویژگی‌های نحوی و مفهومی جملات است. بخش دوم، شامل اختصاص دادن مقدار مناسب به این مکان‌ها می‌باشد.

یکی از روش‌های تولید منحنی گام، برای سیستم تبدیل متن به گفتار، روش تیلت[2] می‌باشد؛ که بر اساس نظریه‌ی واج‌شناسی لایه‌ای پایه‌گذاری شده است و درآن، رویدادهای آهنگین گفتار، به صورت مستقل از هم، در نظر گرفته می‌شوند. مدل تیلت با تولید رویدادهای کانتور گام و اتصال آن‌ها به یکدیگر، کانتور گام[3] را تولید می‌نماید. هر رویداد دارای تعدادی پارامتر می‌باشد که باید تخمین زده شوند.

- دیرش: میزان دیرش واج، تحت تاثیر عوامل مختلفی قرار می‌گیرد. بعضی از عوامل به طور مستقیم و بعضی غیرمستقیم و در تعامل با دیگر عوامل، دیرش را تحت تاثیر قرار می‌دهند. این تاثیرات از خود واج و واج‌های قبلی و بعدی ناشی می‌شوند. همچنین واج‌های انفجاری، با واج‌های طولانی مدت بعدی، همبستگی دارند.
- شدت: یکی از واحدهای اصلی نوای گفتار، انرژی گفتار است که در ارسال مفهوم ذهنی متن، مؤثر است. معمولا گوینده مواضع تکیه و مورد تاکید را با انرژی بیشتری قرائت می‌کند. انرژی دارای نام‌های دیگری مانند شدت و بلندی دامنه نیز می‌باشد.

[1] *Phase*
[2] *Tilt*
[3] *Pitch Contour*

برای مدل‌سازی انرژی از روش منحنی‌های قطعه قطعه، استفاده می‌شود. در این روش، برای هر واج، یک منحنی تولید می‌شود و از اتصال کلیه‌ی این منحنی‌ها، منحنی انرژی برای کل گفتار، به دست می‌آید. برای مدل کردن منحنی هر واج، چندجمله‌ای درجه دوم در نظر گرفته شده و از شبکه‌ی عصبی[1]، ماشین پشتیبان بردار[2]، و مدل مارس[3] برای تخمین ضرایب چندجمله‌ای‌ها استفاده می‌شود.

- درنگ: یکی از مراحل تخمین نوای گفتار در یک سیستم تبدیل متن به گفتار، می‌تواند تخمین موقعیت‌های وجود درنگ در میان کلمات متن ورودی باشد. این پارامتر، تاثیر زیادی بر روی سایر پارامترهای گفتار دارد.
- فاز: به دلیل گسترده بودن دادگان و تفاوت آن‌ها در نوای گفتار، در هنگام اتصال آن‌ها، ناپیوستگی در طیف و عدم تطبیق فاز صورت می‌گیرد. بنابراین، باید بهترین واحدها با کمترین اختلاف فاز، به هم اتصال پیدا کنند. با استفاده از روش‌های تخمین، بهترین گزینه‌ها انتخاب می‌شود و کنارهم چسبانده می‌شود.
- استفاده از تکیه و آهنگ گفتار: باید با بررسی کلمات و جملات به قوانینی دست پیدا کنیم و با برنامه‌نویسی این قواعد را بر آنان حاکم سازیم. تکیه، در جاهای مختلفی از کلمات وجود دارد. مثلا در اسم، صفت، عدد، ضمیر و ... هرکدام به گونه‌ای خاص می‌باشد. مثلا برای اصوات ندا، تکیه در ابتدای آن‌ها قرار دارد و برای برخی، افعال تکیه در انتهایشان واقع می‌شود.

  همچنین، آهنگ جملات نیز بسته به این‌که جمله‌ی مورد نظر؛ خبری، پرسشی، تعجبی و ... باشد، متفاوت خواهد بود.

[1] *Neural Network*
[2] *SVM*
[3] *Mars*

۲-۸. ارزیابی [۱]

ارزیابی، بخشی جدایی‌ناپذیر در دوره‌ی حیات یک محصول است و برای تعیین محصول بهتر، نیاز به ارزیابی محصولات می‌باشد. هدف از مقایسه، به‌کارگیری مجموعه‌ای از تست‌های قابل مدیریت و ساده است که به وسیله‌ی آن، بتوان بازدهی سیستم‌ها را به‌طور کمی و با استفاده از اعداد و ارقام بیان کرد.

راه‌های مختلفی برای تست سیستم متن به گفتار، تا به حال ارائه شده است؛ ولی هنوز یک روش ارزیابی جامع وجود ندارد که بتواند نتایج صحیح را به درستی شناسایی کند.

۲-۸-۱. ارزیابی جنبه‌های متفاوت صوتی

ارزیابی جنبه‌های متفاوت صوتی، به معنی ارزیابی خروجی بخش سنتزکننده گفتار، در یک سیستم تبدیل متن به گفتار است. بدین منظور، تست‌های متعددی قابل انجام است. انواع این تست‌ها عبارتند از: تشخیص قافیه برای تشخیص میزان قابل‌فهم‌بودن اصوات تولید شده، خصوصا همخوان اول، وسط، و آخر کلمات و بعضی تست‌های دیگر که برای ارزیابی نحوه‌ی انتقال از واکه به همخوان، میزان قابل-فهم‌بودن واکه‌ها، خوشه‌های همخوان‌ها، کلمات در جملات، کلمات چندسیلابی و جملات استفاده می‌شوند. این ارزیابی‌ها به طور کلی در دو سطح انجام می‌شوند: ارزیابی قطعه‌ای و ارزیابی در سطح جمله. معروف‌ترین تست در این قسمت، تست تشخیص قافیه[۲] است.

۲-۸-۲. ارزیابی نوا

در این قسمت از شنونده خواسته می‌شود که تعیین کند گفتار حاصله از لحاظ بیان نوا یا احساسات، چقدر مناسب است. یا در نوع دیگری از ارزیابی، می‌توان از کاربر پرسید که فکر می‌کنید جمله شنیده شده؛ چگونه بیان شده است. نوا از نقش مهمی در یک جمله برخوردار است:

---

[۱] مطالب این بخش از [۱۳] گرفته شده است.

[2] *Diagnostic Rhyme Test (DRT)*

- تعیین کلمه یا مجموعه کلماتی که برای کاربر، بیشترین اهمیت را دارد.
- تعیین مرز عبارات و تعیین گروه‌های کلماتی که با هم، یک مفهوم واحد را منتقل می‌کنند.
- تعیین مفهوم و معنای جمله.

۲-۸-۳. ارزیابی خوشایند بودن و طبیعی بودن

در ارزیابی با تست‌های شنیداری، شنونده‌ها با گوش دادن به صدای حاصل از سیستم تبدیل متن به گفتار، نسبت به مواردی چون میزان طبیعی بودن، خوشایند بودن، آزاردهنده بودن و سرعت بیان و مانند آن‌ها امتیاز می‌دهند. میانگین نظرات شنوندگان، بیانگر نتیجه‌ی ارزیابی است. مهمترین بخش در این قسمت، "میانگین امتیازات نظردهی[1]" می‌باشد.

۲-۸-۴. ارزیابی صحت سنتز

صحت سنتز، به معنای صحت بیان عبارات و کلمات موجود در متن است. این نوع ارزیابی، مهمتر از انواع دیگر است. چون باید جمله‌ی خروجی، با متن ورودی هم‌خوانی داشته باشد ! مثلا در یک متن، استثنائاتی شامل اعداد، کلمات خارجی، کوته‌نوشته‌ها، سرنام‌ها، هم‌نویسه‌ها، غلط‌های املایی و ... است. در این ارزیابی، متن‌های مختلفی به سیستم داده می‌شود و توسط کاربر، عبارات صحیح از غلط تشخیص داده می‌شود. سپس با محاسبه‌ی نسبت عبارات صحیح سنتزشده به کل عبارات، می‌توان یک ارزیابی از میزان صحت عملکرد سیستم به دست آورد.

۲-۹. روش‌های طبیعی‌سازی گفتار

در فصل دوم درباره‌ی نوای گفتار توضیح داده شد و پنج پارامتر مهم آن نام برده شد و درباره‌ی هر کدام از آن‌ها توضیح کوتاهی داده شد. فرکانس گام، دیرش، شدت و درنگ پارامترهایی هستند که از

[1] *Mean Openion Score (MOS)*

قدیم تا حال بایستی روی گفتار اعمال می‌شد. برای تخمین هر کدام از آن‌ها در سیستم‌های سنتز امروزی ابزارها و ترفندهای گوناگونی وجود دارد. ابزارها شامل درخت‌های تصمیم‌گیری، مارکوف، شبکه‌های عصبی و ... هستند.

همچنین روش‌های هم‌گذاری زیادی نیز وجود دارد که با استفاده از مدل‌های ریاضی و ابزارهای مختلف قابل پیاده‌سازی می‌باشند.

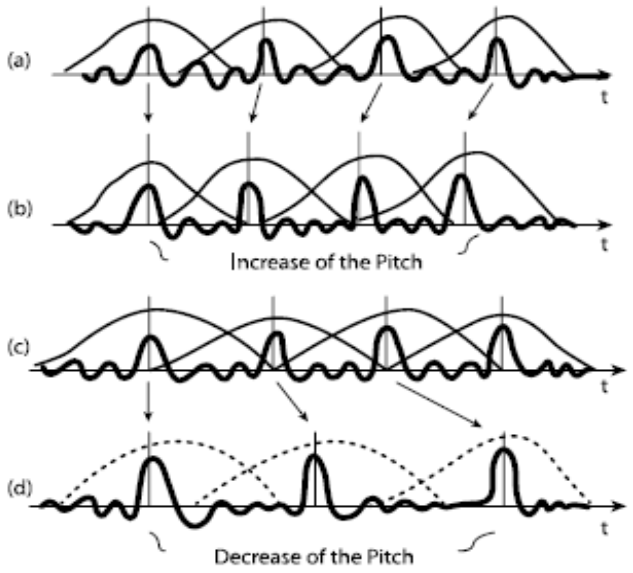
۲-۹-۱. روش‌های بازسازی به روش هم‌گذاری SOLA

روش‌های تحت عنوان SOLA[1] که برای هم‌زمانی و هم‌پوشانی هنگام جمع دادگان، استفاده می‌شوند به دو دسته‌ی هم‌زمان با گام (PSOLA[2]) و هم‌زمان با شکل موج (WSOLA[3]) تقسیم می‌شوند. در بین روش‌های هم‌گذاری PSOLA با انواع FD-PSOLA[4]، TD-PSOLA[5] و MBR-PSOLA[6] روش TD-PSOLA بیشتر از همه برای بازسازی گفتار فارسی مورد استفاده قرار گرفته است.

روش TD-PSOLA تغییرات زیر و بمی و دیرش در حوزه‌ی زمان انجام می‌پذیرد و بنابراین به پردازش اضافی احتیاجی نخواهد بود. البته یافتن نشان‌گرهای گام در این روش الزامی است که برای این کار از روش یافتن زمان‌های بسته شدن حنجره توسط تغییر در مشخصات ایستان طیفی زمان کوتاه قاب‌ها استفاده می‌شود. این روش برای تغییر دادن دیرش و گام از حذف و تکرار قاب‌ها و تغییر میزان هم‌پوشانی استفاده می‌کند که در شکل ۱ به صورت شهودی نشان داده شده است.

---

[1] *Synchronous OverLap and Add*
[2] *Pitch Synchronous OverLap and Add*
[3] *Waveform Synchronous OverLap and Add*
[4] *Fourier-Domain Pitch Synchronous OverLap and Add*
[5] *Time-Domain Pitch Synchronous OverLap and Add*
[6] *Multi-Band Resynthesis Pitch Synchronous OverLap and Add*



شکل ۲-۲: تغییر بسامد گام با کمک روش PSOLA

در [۱۴] از روش هم‌گذاری PSOLA نیز استفاده شده است.

۲-۹-۲ روش بازسازی مبتنی بر انتخاب واحد

این روش با نام‌های دیگر بازسازی به کمک پیکره‌ی بزرگ، یا بازسازی به روش دنبال‌هم‌چینی مجدد نیز، شناخته می‌شود. در این روش به جای استفاده از دادگان کوچک یا دیکشنری واحدهای دوآوایی، از یک دادگان بزرگ یک تا چند ساعت گفتار، استفاده می‌شود. سعی بر این است که دادگان مورد استفاده در بازسازی را به نحو موثری بزرگ کرد. در این حالت از دادگان با واحدهای صوتی با طول متغیر و نمونه‌های صوتی مختلف که به بهترین نحو با گویش تطابق دارند، استفاده می‌شود. انجام عملیات پردازش سیگنالی روی واحدها باعث کاهش کیفیت آن‌ها و در کل، کاهش میزان طبیعی بودن خروجی سیستم می‌شود. درحالی که در سیستم‌های مبتنی بر انتخاب واحد، مساله‌ی مهم، انتخاب واحدها و نمونه‌های مناسب صوتی از میان تعداد زیادی از واحدها و نمونه‌های صوتی می‌باشد. نکته‌ی دیگر این روش، امکان استفاده از واحدهای بزرگ چون هجاها و کلمات است. به دلیل ظرفیت

و قابلیت زیاد، این روش در دهه‌ی اخیر به مقدار زیادی مورد توجه قرار گرفته است. یک نمونه از این سیستم، CHART می‌باشد. دو نمونه از سیستم نشر مبتنی بر انتخاب واحد برای زبان فارسی در [۱۵] و [۱۶] طراحی و پیاده‌سازی شده است.

در این‌گونه سیستم‌ها مساله‌ی مهم، انتخاب واحدهای صوتی مناسب و سپس نمونه‌های صوتی مناسب از هر واحد می‌باشد. در مرحله‌ی اول سعی می‌شود با کمک یک تابع هزینه، توالی مناسب واحدهای صوتی پیدا شود. این کار با کمک الگوریتم‌های جستجوی پویا همچون الگوریتم جستجوی ویتربی انجام می‌شود. در واقع در این حالت نمونه‌های صوتی یا واحدهای صوتی گره‌های یک گراف درنظر گرفته می‌شوند. بعد از پیدا کردن توالی مناسب واحدهای صوتی که کمترین هزینه را دارند، اکنون بین نمونه‌های هر واحد صوتی انتخاب شده جستجو انجام می‌شود و سعی می‌شود بهترین نمونه که به وسیله‌ی یک تابع هزینه، مشخص می‌شود، پیدا شود. در نهایت با اتصال نمونه‌های پیدا شده، خروجی بازساز ساخته می‌شود. می‌توانیم مرحله‌ی اول الگوریتم انتخاب واحد که میان انواع واحدهای صوتی جستجو می‌کند را انتخاب واحد، و مرحله دوم که میان نمونه‌های صوتی یک واحد جستجو می‌کند را انتخاب نمونه یا قطعه بنامیم.

۲-۹-۳. روش هم‌گذاری واحدهای حساس به بافت نوایی

در مقاله [۱۴]، سه مدل مختلف را با هم مقایسه کرده است. در یک حالت از هجاهایی استفاده شده که هرکدام به تحقیق نگارنده‌اش در جمله نقشی می‌گیرد و نوای آن نسبت به بقیه متفاوت است. در این‌جا متن به سه گروه آهنگی IP، زیروبمی AP و کلمه واجی PW تقسیم‌بندی شده و هر کدام از آن-ها حالت آغازی و پایانی دارند. این برابرهای آوایی وابسته به بافت نوایی را هجاهای نوایی نامیده است که تعداد آن‌ها شش عدد می‌باشد. در حالت دوم، از واحدهای بدون نوا استفاده شده و توسط ابزارهای مختلف دیرش، شدت و فرکانس گامشان تصحیح می‌شود و اصطلاحا روی آن‌ها پردازش انجام می-شود. در نوع سوم نیز از واحدهای ساده بدون هیچ پردازشی برای سنتز استفاده شده است. نتایج

تحقیق نشان داده است که روش اول و دوم و بین آن‌ها روش اول پاسخ بسیار خوبی دارد و گفتار خروجی‌اش طبیعی‌تر بوده است. این تحقیق می‌خواهد ثابت کند که استفاده از نوع طبیعی دادگان بهتر از مدل‌سازی جایگاه نوایی می‌باشد.

فصل سوم:

مبانی تئوری

۳-۱. مقدمه

در این فصل، درباره‌ی مبانی تئوری مورد نیاز برای انجام پایان‌نامه مطالبی آورده‌ایم. مبانی تئوری مورد نیاز برای انجام کار، به دو دسته‌ی عمده تقسیم می‌شود: ۱- زبان‌شناسی و بررسی تکیه، مکث و آهنگ جملات ۲- بررسی انواع دادگان و روش‌های استخراج، برای یک سیستم متن به گفتار

در بخش مبانی تئوری زبان‌شناسی، ابتدا درباره‌ی پردازش متن و چالش‌هایی که در آن به وجود می‌آید به طور مختصر توضیح داده شده و سپس به صورت تخصصی‌تر درباره‌ی نقش کلمات در جمله، و چگونگی شناسایی آن‌ها به وسیله ابزارهایی مانند درخت تصمیم‌گیری و مدل مخفی مارکوف بحث شده است.

جملات به اشکال مختلفی تقسیم‌بندی شده‌اند: خبری، سوالی، تعجبی، حزن‌انگیز، پرسشی، خشمگین، بیان‌گر حالت و ... . برای این اشکال نیز چند شکل جامع‌تر برگزیده شده است که تقریبا بیشتر جملات فارسی را پوشش می‌دهند. در فصل روش پیشنهادی، تقسیم‌بندی نگارنده و دلایل انتخابش، به تفصیل بیان شده است.

منظور ما از آوردن این قسمت در پایان‌نامه این است که بگوییم پردازش متن، یک پروسه‌ی بزرگ و طولانی است. تا آنجا که متن به جملات خود و کلمات به هجاها و دایفون‌های تشکیل‌دهنده‌اش تقسیم شود، باید پیش رفت؛ سپس در قسمت سنتز گفتار، از چسباندن دایفون‌های مناسب در جایگاه خودشان، جملات با آهنگ و تکیه‌ی واژگانی مطلوب، تولید می‌شوند.

در بخش مبانی تئوری مربوط به سنتز گفتار، می‌خواهیم راجع به کارهای انجام شده توسط محققان، در زمینه‌ی دادگان گفتاری و متنی و روش‌های استخراج واحدهای گفتاری مطالبی را بیان کنیم. در نهایت تصمیم نگارنده بر این خواهد بود که روش استخراج واحدهای گفتاری خود را انتخاب کند: استخراج دایفون‌ها با استفاده از سه روش موازی شنیداری، دیدن شکل موج گفتار و طیف‌نگاشت گفتار.

۳-۲. پردازش متن (تعیین نقش کلمه، نوع جمله، مرز هجایی و دایفونی)

در قسمت پردازش متن، باید متن‌ها به جملات تشکیل‌دهنده‌شان تقسیم شوند. جملات را می‌توان از نظرهای مختلفی تقسیم‌بندی کرد. تقسیم‌بندی مورد نظر ما در این پایان‌نامه، بر اساس حالت بیانی جمله است؛ که در قسمت روش پیشنهادی، تقسیم‌بندی خود را بیان می‌کنیم و دلایل آن را ذکر می‌کنیم. یکی از روش‌های ساده برای این کار، مشخص بودن علائم آخر جمله مثل "."، "؟"، "!"و "؟!" است. علاوه براین علائم ظاهری می‌توان از علائم دیگر که اکثرا دستوری هستند، استفاده کرد. مثلا جمله‌ای که با "آیا" شروع می‌شود و یا جمله‌ای که حرف "چه" در ترکیب آن است.

در ادامه‌ی پردازش متن، به قسمت جدا کردن کلمات تشکیل‌دهنده و مشخص کردن نقش هر کدام از آن‌ها در جمله می‌پردازند. دانستن نقش کلمه، به ما کمک می‌کند که با استفاده از اطلاعات دستوری، در ادامه‌ی کار، مکان تکیه‌ی کلمات را نیز مشخص کنیم. اطلاع از مکان تکیه، در بخش سنتز، مورد نیاز است؛ چون در دادگان، هجای تکیه‌دار و هجای بدون تکیه هر کدام دارای یک داده‌ی مجزا می‌باشد. مثلا عبارت "دست نیاز" در جمله‌ای آمده است. باید بدانیم که این عبارت یک ترکیب وصفی است و در یک ترکیب وصفی جایگاه تکیه در موصوف و صفت، از لحاظ دستوری به چه گونه است. این اطلاعات با ابزارهای مختلفی قابل شناسایی می‌باشند. برای مثال در [۱۰] روش‌های مختلفی برای این کار ذکر شده است. یکی از این روش‌ها، استفاده از الگوی مخفی مارکوف به کمک پیکره‌ی بی‌جن‌خان است. به این ترتیب که ابتدا تمام نقش‌های موجود برای یک واژه از پیکره، فراخوانی می‌شود؛ سپس با توجه به کلمات قبل و بعد[۱] و استفاده از احتمالات شرطی، نقش کلمه مشخص می‌شود. [۱۱] همچنین در روشی دیگر از ابزار درخت‌های تصمیم‌گیری و پیکره‌ی بی‌جن‌خان استفاده شده است.

[۱] یکی از پارامترهایی که در مدل مخفی مارکوف اهمیت دارد، عمق حافظه می‌باشد. یعنی تشخیص نقش کلمه، با استفاده از چند کلمه‌ی قبل و چند کلمه‌ی بعد. با توجه به تعداد این کلمات، عمق حافظه مشخص می‌گردد.

پایگاه داده‌ی دیگری، برای ذخیره‌سازی دایفون‌ها و مرز هجاها در کلمات مختلف، مورد نیـاز اسـت. در این‌جا می‌توان پیکره‌ای شبیه به پیکره‌ی بی‌جن‌خان تولید کرد؛ با این تفاوت کـه در پیکـره‌ی بـی-جن‌خان نقش دستوری کلمات در جمله مشخص شده است ولی در این‌جا بـرای هـر کلمـه‌ای کـه در پیکره موجود است، مرز دایفون‌ها و هجاها مشخص می‌شود.

از آن‌جا که شرح و بسط این موضوع در این پایـان‌نامـه لازم نیسـت، کسـب اطلاعـات بیشـتر را بـه عهده‌ی خواننده می‌گذاریم.

۳-۳. مشکلات موجود در پردازش متن

در پردازش متون زبان طبیعی با زبان نوشتاری سروکار داریـم. در ایـن مسـئله، اگـر چـه، بـه جهـت ازدست‌دادن اطلاعات گویشی، مانند لحن گوینده، آهنگ صدا، تاکید و مکث، بـا مشـکلات و ابهامـاتی مواجه می‌شویم، ولی در مقابل با شکل محدودتری از زبان، کار می‌کنیم . بسـیاری از بی‌ترتیبی‌هـای زبان، متعلق به زبان گفتاری است و در زبان نوشتاری، بیشتر قالب‌های دستوری رعایت می‌شوند و لذا تهیه‌ی دستور زبان پوشاننده‌ی تمام متن، سـاده‌تر اسـت. بـرای مثـال در تـلاش بـرای سـاخت یـک سامانه‌ی پردازش و درک متون مانند زبان فارسی با مسائل و مشکلاتی مواجه می‌شویم کـه بعضـی از آن‌ها، در بیشتر زبان‌ها بروز کرده و برخی خاص زبـان فارسـی می‌باشـند . هـم‌چنـین برخـی از ایـن پیچیدگی‌ها به طبیعت زبان و نارسـایی‌های قواعـد زبان‌شناسـی مربـوط و برخـی دیگـر برخاسـته از مشکلات ایجاد سامانه‌های هوش مصنوعی است . برخـی از مشـکلات بـه شـرح ذیـل هسـتند[۶] (در پیوست توضیح کامل‌تری درباره‌ی این موارد داده شده است):

- مواجهه با چندمعنایی و چندنقشی بودن کلمات
- حذف کلمات و عبارات به قرینه‌ی لفظی یا معنوی
- استفاده از افعال مرکب، اصطلاحات و ضرب‌المثل‌ها
- تعیین طبقه‌ی اسامی

- بی‌ترتیب‌بودن واحدهای زبان
- کسره‌ی اضافه و حذف آن
- عدم تطابق اجزای جمله
- وجود ساختار جملات یکسان با معانی و نقش‌های متفاوت
- مشکلات ناشی از ابهام زبان طبیعی

۳-۴. روش‌های تخمین پارامترها برای استفاده در سیستم متن به گفتار

در یک سیستم تبدیل متن به گفتار لازم است برای تخمین مقادیر و پیدا کردن جواب مناسب، از ابزارهایی استفاده کرد. به همین ترتیب، می‌توان مقادیر مربوط به پارامترهایی را که در بخش نوای گفتار ذکر شده، به گفتار مصنوعی، اضافه کرد؛ یعنی مقادیر گام، دیرش، شدت، درنگ و فاز. برای تعیین هر کدام از آن‌ها، از روش‌های متفاوتی می‌توان استفاده کرد؛ زیرا روش‌های مختلف ارائه شده، بر روی داده‌هایی با مشخصات خاص، عملکرد درستی از خود نشان می‌دهند. این روش‌ها را، در این‌جا می‌توان به دو دسته‌ی کلی قانون‌گرا[1] و داده‌گرا[2] تقسیم کرد. در روش‌های مبتنی بر قانون، سیستم بر اساس یک سری قوانین از پیش تعیین شده، عمل می‌کند. معروف‌ترین مدل قانون‌گرا، مدل کلات[3] می‌باشد. با توجه به این که نوشتن کلیه‌ی قوانین، در عمل بسیار سخت است و در برخی موارد، قانون خاصی که برای همه‌ی موارد بتوان تعمیم داد، پیدا نمی‌شود؛ لذا روش‌های داده‌گرا، جای روش‌های قانون‌گرا را گرفته‌اند.

در مدل‌های داده‌گرا، ساختار مدل، به طور خودکار از رویدادهای آموزشی به دست می‌آید. از جمله این روش‌ها؛ می‌توان به آنتروپی، درخت تصمیم‌گیری، شبکه‌ی عصبی، مدل مارس، ماشین پشتیبان

[1] *Rule-Based*
[2] *Data-Based*
[3] *Klatt*

بردار، مدل مخفی مارکوف و ... اشاره کرد. در پیوست، توضیح بسیار مختصری درباره‌ی روش‌های نام-برده داده شده است.

۳-۵. تکیه، آهنگ، مکث و کاربردشان در طبیعی‌سازی سنتز گفتار

این پارامترها در گفتار فارسی باعث می‌شوند، افراد منظور سخن خود را، بهتر به گوش شنونده برسانند. برای این منظور، دو پدیده‌ی نوایی یعنی تکیه در کلمات، و آهنگ در جملات را مورد بررسی قرار می‌دهیم. پدیده‌ی مکث نیز باید به صورت قانون‌مند در سنتز استفاده شود (ولی مکث آن‌چنان قانون‌مند نیست). برای بررسی این پدیده‌ها، باید از دانش زبان استفاده کرد. سپس برای پیاده‌سازی آن‌ها باید از دانش برنامه‌نویسی و ابزار موجود در رایانه استفاده کرد. [۱۷]

۳-۵-۱. تکیه

"تکیه" وسیله‌ای است که "واژه" را در جمله برجسته می‌سازد. تکیه در یک کلمه، معمولا با اعمال هم‌زمان تغییر درجه‌ی زیر و بمی، تغییر دیرش و شدت‌دار کردن هجای مناسب واژه‌ی تکیه‌دار مشخص می‌شود. تکیه در کلمه، گاهی اوقات تعیین‌کننده‌ی معنی آن است. مثلا در کلمه‌ی "ولی" اگر تکیه در ابتدای آن باشد به معنی "اما"، و اگر در انتهای آن باشد، به معنی "سرپرست" می‌باشد. بر اساس تحقیقات دکتر ساسان سپنتا، دیرش در هجای تکیه‌بر، ۰/۰۸ ثانیه بیشتر از واکه‌های بدون تکیه می‌باشد؛ یعنی دیرش هجای تکیه‌بر، دو برابر دیرش هجای بدون تکیه است.

در زبان فارسی، کلمات در جمله با توجه به نقششان، تاکید خاصی می‌پذیرند. معمولا تکیه‌ی کلمات در آخرین هجای آن‌هاست. ولی وقتی در جمله قرار می‌گیرند، شاید تکیه‌ی آن‌ها حذف، اضافه یا جابه‌جا شود. یک کلمه در جمله، می‌تواند نقش‌هایی مانند اسم (جمع، مفرد؛ نکره، معرفه؛ عام، خاص؛ و ...)، صفت (ترتیبی، شمارشی، عالی، تفضیلی، و ...) و ... را بپذیرد، برای مثال، تکیه در اصوات

ندا، در ابتدا و در بیشتر اسامی، در انتهای آن‌ها قرار می‌گیرند. در فصل روش پیشنهادی، تاکیددار یا بدون تاکید بودن هجاها، باعث می‌شود، کلمات با تکیه‌ی صحیح، در سنتز خروجی تولید شوند.

انواع واژگان در زبان فارسی تکیه‌ی خاص خود را می‌پذیرند. همچنین وقتی این واژگان در قالب جمله قرار می‌گیرند، ممکن است مدل تکیه‌ی آنها عوض شود. انواع واژگان در فارسی تقسیم‌بندی کلی روبرو را دارد: اسم، صفت، فعل، قید، عدد، حرف اضافه و ادات ربط. در [۱۸] به طور مفصل درباره‌ی تکیه‌ی کلمات در جملات، توضیح داده شده است.

۳-۵-۲. آهنگ

آهنگ، همچون روح زبان است که معمولا به جمله اعمال می‌شود و بیانگر حالت عاطفی و احساسی گوینده می‌باشد. مثلا فعل "تو آمدی" را می توان با آهنگ‌های مختلفی از قبیل خبری، پرسشی، تعجبی، پرسشی-تعجبی و ... بیان کرد. البته هر کدام از این مدل‌های جامع آهنگ جملات، زیر-مجموعه‌های زیادی دارند؛ مثلا جمله‌های خبری شامل التماسی، تمسخری، تعارفی و ... می‌باشد.

بحث آهنگ در جمله، بسیار پیچیده و طولانی است که برای تحقیق مفصل آن باید به [۱۸] مراجعه کرد. با خواندن مباحث زبان‌شناسی و جمع‌بندی آن‌ها، می‌توان به چند حالت کلی دست پیدا کرد که بیشتر جملات را پوشش دهند. در این‌جا به چند حالت مختلف جملات با آهنگ‌های متفاوت از [۱۸] اشاره می‌شود. نگارنده در فصل روش پیشنهادی، درباره‌ی تقسیم‌بندی زبان فارسی به آهنگ-های تشکیل‌دهنده‌اش توضیح داده است.

۳-۵-۳. مکث

در [۱۸] درباره‌ی مکث نوشته شده:

"کلام، رشته‌ای پیوسته نیست؛ بلکه در آن درنگ هست. این مکث یا درنگ، معمولا به سبب بست-واج‌هاست، یا به ضرورت تنفس، یا برای روشن‌تر ساختن معنی واژه‌ها. بنابراین، درنگ را می‌توان به چند دسته تقسیم کرد:

۱- درنگ میان بست‌واج‌ها: به علت بسته شدن مجرای تنفس به وجود می‌آید و مورد بحث ما نیست.

۲- درنگ میان تکواژها: در بعضی زبان‌ها، بین دو تکواژ یک واژه، ممکن است درنگی باشد که ایجاد تمایز معنایی کند (بدون تفاوت واجی و تکیه‌ای و آهنگی)، که تنها وجود درنگ، موجب تفاوت معنایی می‌شود.

۳- درنگ میان واژه‌ها: میان دو واژه، درنگ، بالقوه میسر است.

۴- درنگ بعد از جمله‌ی کامل یا بزرگ‌ترین قسمت از یک جمله‌ی طولانی، که عملا در جایی صورت می‌گیرد که از نظر معنایی مجاز باشد. "

به عقیده‌ی نویسنده [۱۸]، درنگ تمایز دهنده معنایی بین تکواژهای زبان فارسی وجود ندارد و درنگ در قالب‌های کلمه و عبارت و جمله و ... می‌تواند وجود داشته باشد. از نظر عملی، وقتی کسره‌ی اضافه بین کلمات نباشد، از مکث استفاده می‌شود و در صورت وجود "،" مکث اجباری است.

۳-۶. دادگان‌های متنی و گفتاری مورد استفاده در تبدیل متن به گفتار

یکی دیگر از ملزومات در سیستم‌های تبدیل متن به گفتار، داده‌های زبانی می‌باشد. در یک سیستم تبدیل متن به گفتار، حداقل به سه نوع دادگان و داده‌ی زبانی نیاز داریم. این دادگان‌ها شامل واژگان، دادگان‌های گفتاری و پیکره‌های متنی می‌باشند. [۱۱]

در مورد پیکره‌های متنی در [۲] اینگونه آمده است:

"پیکره‌های متنی به منظور تحلیل‌های آماری، صحت‌سنجی فرضیه‌ها و بررسی رخداد یا صحت قواعد زبانی در حوزه‌ای مشخص به کار می‌روند. پیکره‌های متنی به عنوان پایگاه دانش اصلی در

زبان‌شناسی رایانه‌ای، خصوصا در حوزه‌ی خط و زبان، شناخته می‌شوند. پیکره‌های متنی می‌توانند تک‌زبانی، شامل متون با یک زبان، و یا چندزبانی، شامل متون با چندین زبان باشند، که مورد اخیر معمولا برای مقایسه‌ی نظیر به نظیر، ساختاردهی می‌شود و پیکره‌ی موازی-منطبق نامیده می‌شود."

دادگان گفتاری به دو دسته تقسیم می‌شود:

۱- متون خوانده شده: قطعاتی از یک کتاب، گزارش خبری، لیست کلمات، دنباله‌ی اعداد.

۲- گفتار طبیعی: جلسات و گفتگوها، نقل داستان.

برخی از فواید استفاده از دادگان گفتاری، عبارت‌اند از: صرفه‌جویی در زمان برای دستیابی به داده-های گفتاری، دستیابی به حجم بالایی از گفتار، و قابلیت جستجو در صورتی که این امکان برای دادگان فراهم شده باشد.

دادگان گفتاری، از منظر نوع گفتار ذخیره شده به دو دسته تقسیم می‌شوند:

۱- ادای جملات مختلف مثل فارس‌دات که در بازشناسی گفتار کاربرد دارد.

۲- واحدهای آوایی خاص که در تبدیل متن به گفتار کاربرد دارند.

۳-۷. ویژگی‌های دادگان‌های گفتاری مورد استفاده در سنتز گفتار

در گردآوری دادگان‌های گفتاری برای یک سیستم سنتز گفتار، نکات زیر می‌بایست مورد توجه قرار گیرند:

۱- گوینده‌ی انتخاب شده، صدای مناسب داشته باشد.

۲- گفتار به لحاظ آوایی، غنی باشد و تکرارهای کافی و پوشش مناسبی از واحدهای گفتاری مورد نیاز را، داشته باشیم.

۳- پایگاه داده به اندازه‌ی کافی بزرگ باشد، اما ضبط آن در تعداد جلسات کمی انجام شود؛ تا یکسانی داده‌ها حفظ شود.

۴- در روش بازسازی مبتنی بر انتخاب واحد از دادگان بزرگ، دادگان گفتاری بهتر است، شامل؛ جملات جداگانه، نوشت‌پاره و نمونه‌هایی از مکالمه باشد. از جملات نقل قولی زیاد استفاده شود. گفتار شامل کلمات پیوسته باشد نه کلمات گسسته، چرا که در غیر این صورت؛ گفتار نهایی کاملاً غیر طبیعی خواهد بود.

۵- فرکانس نمونه‌برداری مناسب باشد. فرکانس ۱۱۰۲۵ هرتز و یا ۱۶۰۰۰ هرتز مناسب است.

۶- میکروفون از کیفیت مناسبی برخوردار باشد. میکروفون‌های دینامیکی با کیفیت، برای این کار مناسب هستند.

۷- فاصله‌ی دهان از میکروفون همواره ثابت باشد؛ لذا می‌توان از میکروفون نصب شده روی هدفون، استفاده کرد.

۸- محیط ضبط صدا، ساکت باشد.

۹- بسته به واحدهای گفتاری مورد نیاز، دادگان به طور دقیق برچسب‌گذاری شود.

۱۰- برای استفاده در مدل‌سازی نوای گفتار، جملات، هجاها، واج ها و خصوصیات زبرزنجیری گفتار نیز، برچسب‌گذاری شوند.

۱۱- در روش بازسازی مبتنی بر انتخاب واحد از دادگان بزرگ، دادگان حاوی اعداد و تاریخ، اسامی خاص و همچنین الگوهای نوایی متفاوت باشد.

یک پایگاه داده از کلمات جداگانه می‌تواند مبنایی برای ایجاد گفتار به روش هم‌گذاری، قرارگیرد؛ ولی گفتار تولید شده با این روش، به شدت غیر طبیعی خواهد بود. بنابراین، اگر بخواهیم که سیستم بازسازی گفتار، بتواند داستان‌های خبری را بخواند، بهتر است که پایگاه داده نیز، حاوی همان نوع داده باشد. علاوه بر پوشش مناسب آواهای زبان، پوشش ویژگی‌های نوایی نیز بسیار مهم است. اگر هدف این باشد که سیستم بازسازی گفتار، بتواند دامنه‌ی وسیعی از لحن‌ها را ایجاد کند، باید به تعداد کافی از هر یک از آن‌ها را داشته باشیم تا بتوانیم مدل‌های تولید نوا را با استفاده از آن‌ها آموزش دهیم.

۳-۸. دادگان‌های فارسی

۳-۸-۱. دادگان گفتاری فارس‌دات

این پایگاه داده جهت پیاده‌سازی سیستم‌های آزمایشگاهی بازشناسی گوینده و یا بازشناسی گفتار مستقل از گوینده، با واژگان کوچک و متوسط کمتر از ۱۰۰۰ کلمه و برای پژوهش‌های موردی و پایان‌نامه‌های دانشجویی مناسب می‌باشد. گفتار هر گوینده، فقط حدود ۱۰۰ ثانیه است و کمتر از ۱۰٪ واج‌گونه‌های زبان را شامل می‌شود، بنابراین؛ برای تبدیل متن به گفتار مناسب نمی‌باشد. این دادگان دارای جملات مجزا و کلمات گسسته می‌باشد، لذا برای بررسی‌های نوایی در تبدیل متن به گفتار، کمتر استفاده می‌شود. مشخصات کامل این دادگان در [۴] آمده است.

۳-۸-۲. دادگان فارس‌دات بزرگ

ویژگی‌های فارس‌دات بزرگ عبارت‌اند از:

- تعداد گویشوران: ۱۰۰ نفر.
- انواع گویش‌ها: گویش‌های رایج: تهرانی، ترکی، اصفهانی، جنوبی، شمالی، خراسانی، بلوچی، کردی، لری و یزدی.
- تنوع سن و تحصیلات گویندگان.
- محیط ضبط: اتاق معمولی (نسبتاً آرام).
- انواع میکروفون: میکروفون رومیزی تک جهته، دو میکروفون یقه‌ای، و یک میکروفون هدفونی.
- ضبط تغییرات ارتعاش تارهای صوتی با کمک دستگاه لارینژوگراف به همراه گردنبد حس‌گر[1].
- متن دادگان به ازای هر گوینده: ۲۰ الی ۲۵ صفحه مطلب نوشتاری فارسی.

[1] استفاده می‌شود، و از لوازم این پردازش‌ها، معین بودن محل (Pitch) در تعدادی از روش‌های بازسازی گفتار، از پردازش‌های هم‌زمان با گام آغازه‌های دوره‌ی تناوت گام می‌باشد. از این رو، ضبط سیگنال به صورت دوکاناله انجام می‌پذیرد که یکی از سیگنال‌ها گفتار و دیگری سیگنال حنجره (تکانه‌های تارآوا) است.

- موضوعات متون: فرهنگی، اقتصادی، اجتماعی، علمی و... .
- حجم دادگان: ۲۲ گیگا بایت.
- فرکانس نمونه‌برداری: ۲۲/۰۵ کیلوهرتز.
- نسبت سیگنال به نویز: ۲۸ دسی بل.
- ۱۴۰ ساعت گفتار تک‌کانال (حدود ۲۸۰ ساعت گفتار دوکانال از چهار میکروفون مختلف).
- تقطیع: در سطح کلمه.

۳-۸-۳. دادگان هجائی پژوهشکده پردازش هوشمند علائم

دارای مشخصات زیر است:

- دادگان شامل صدای یک گوینده‌ی مرد و حاوی ۶۰۰۰ هجای پرکاربرد فارسی.
- بخش بزرگ هجاها از صورت واج‌نویسی شده فارس‌دات بزرگ، و بخش دیگر از فرهنگ معین.
- ضبط تغییرات ارتعاش تارهای صوتی با کمک دستگاه لارینژوگراف به همراه گردنبد حس‌گر.
- نمونه‌برداری با فرکاس ۲۲۰۵۰ هرتز و دقت ۱۶ بیت به ازای هر نمونه، در اتاق ساکت و میکروفون دینامیکی.
- تقطیع گفتار با نرم‌افزار WaveStudio ، عدم پیوستگی و اعوجاج طیفی حداقل، محل برش در ابتدای پریود و در زمان عبور از صفر سیگنال.
- اضافه شدن فرآیند واژ-واجی به دادگان.

۴-۸-۳. دادگان دایفونی پژوهشکده پردازش هوشمند علائم

دارای مشخصات زیر می‌باشد:

- دادگان شامل صدای دو گوینده‌ی مرد و زن و حاوی ۹۶۶ دوآوایی.

- برای هر یک از دوآوایی‌ها، ۳ کلمه که دارای بیشترین بسامد بوده و حاوی دوآوایی مورد نظر نیز بوده است، انتخاب شده و برای دوآوایی‌هایی که در واژگان موجود نبوده، از ترکیب کلمات و یا استفاده از کلمات بی معنی کمک گرفته شده است.
- هیچ دوآوایی از داخل کلمات تک‌هجایی، استخراج نشده است.
- ضبط تغییرات ارتعاش تارهای صوتی با کمک دستگاه لارینژوگراف.
- نمونه‌برداری با فرکاس ۲۲۰۵۰ هرتز و دقت ۱۶ بیت به‌ازای هر نمونه، در اتاق ساکت و میکروفون دینامیکی.

۳-۸-۵. دادگان گفتاری تهیه شده در آزمایشگاه پردازش هوشمند سیگنال و گفتار دانشکده مهندسی کامپیوتر و فناوری اطلاعات دانشگاه صنعتی امیرکبیر

به منظور استفاده در تبدیل متن به گفتار، طراحی و ضبط شده است.

- مشخصات گویشور: مرد ۴۰ ساله.
- نوع گویش: گویش تهرانی.
- میزان تحصیلات: لیسانس.
- محیط ضبط: اتاق ساکت.
- نوع میکروفون: میکروفون رومیزی تک جهته دینامیکی با کیفیت بالا.
- متن دادگان: ۴۰۰ صفحه مطلب شامل:
  - جملاتی از فارس‌دات کوچک غیر تلفنی که شامل ۴۰۰ جمله کوتاه است که به اشکال سوالی، خبری، امری و تعجبی مورد استفاده قرار گرفته است.
  - جملاتی از فارس‌دات بزرگ که با توجه به میزان تنوع جملات موجود در متن، انتخاب شده است. متن‌ها شامل موضوعات مختلف از قبیل طنز، اقتصادی، فرهنگی،

سیاسی، جنگ، ورزشی، صنایع، پزشکی، هنری، مذهبی، آگهی، داستان و خـاطره و... می‌باشند.

- یک متن نمایش‌نامه در حدود ۱۳ صفحه نیز، مورد استفاده قرار گرفته است که علت انتخاب آن، ایجاد فراز و نشیب‌های نوایی کافی در دادگان گفتار بازسازی می‌باشد.

- حجم دادگان: ۱۱ ساعت.
- فرکانس نمونه برداری: ۱۶۰۰۰ هرتز.
- نسبت سیگنال به نویز: بالاتر از ۳۴ دسی بل.
- تقطیع: تقطیع حدود ۲ ساعت از این دادگان در سطح واج به طور خودکار و سپس کنتـرل آن به صورت دستی.

علاوه بر دادگان گفتاری فوق، دو دادگان گفتاری جدید بـا صـدای زن نیـز، در آزمایشـگاه پـردازش هوشمند سیگنال و گفتار دانشگاه صنعتی امیرکبیر، در حال تهیه است؛ که یکی از آن‌ها دادگانی است که برای بازسازی مبتنی بر دوآوایی و دنباله‌های آوایی پرکاربرد است، و دیگری نیـز دادگـان گفتـاری جهت استفاده در روش بازسازی مبتنی بر انتخاب واحـد اسـت کـه بـرای بازسـازی مبتنـی بـر سـایر واحدهای آوایی همچون دوآوایی و واج نیز، قابل استفاده است. در تهیه‌ی ایـن دادگـان‌هـا اسـتفاده از گویندگان باتجربه و خوش‌صدا، از جمله گویندگان صداوسیما مورد نظر بـوده اسـت، و ضـبط صـدا در استودیوی ضبط، و با تجهیزات مناسب انجام شده است.

۳-۹. پیکره‌های متنی کارشده در ایران

- پیکره‌ی متنی زبان فارسی پژوهشکده پردازش هوشمند علائم.
- پیکره‌ی بی‌جن‌خان (بخشی از پیکره‌ی پژوهشگر پردازش هوشمند علائم).
- پیکره‌ی پروژه‌ی شیراز.

- پیکره‌ی فارسیFLDB [1].
- پیکره‌های مجموعه‌ی همشهری.
- پیکره‌ی فارسی تحلیل احساس سِنتی‌پِرس
- پی.سی.ای.سی ۲۰۰۸ (پیکره‌ی مرجع ضمیر)
- ... .

۳-۱۰. روش‌های تقطیع واحدهای گفتار

در این بخش، مقالات خوانده شده را بررسی می‌کنیم. می‌خواهیم روش‌های مختلف تقطیع واحد که در مقالات آمده‌اند را بیان کنیم و در انتها یک روش را برای پایان‌نامه برگزینیم.

۳-۱۰-۱. استخراج هجا

در [۲] این‌گونه آمده است:

- تمام هجاها بر اساس نظرسنجی، از هجای بی‌تکیه از اول کلمه‌ی مجزا، مانند "پَ" در ابتدای کلمه‌ی "پذیرفته".
- ضبط در اتاق آکوستیک و تقطیع و برچسب‌دهی هجایی، و قرارگرفتن در دادگان صوتی هجاها.
- شیوه‌ی نرم‌افزاری آن: ۶۰۰۰ کلمه، که بعضی از آن‌ها بی‌معنی هستند، توسط گویشور مذکر با لهجه‌ی فارسی تهرانی قرائت می‌شوند، و ضبط می‌گردد.
- محل تقطیع به گونه‌ای است که عدم پیوستگی و اعوجاج طیفی بین قطعات متوالی، به حداقل می‌رسد. برای کاهش میزان ناپیوستگی، سعی شده که محل تقطیع، در جاهایی باشد که دامنه‌ی شکل موج، چندان زیاد نیست. کم بودن دامنه در مرزها، تا حدودی، ناپیوستگی

[1] *Farsi Linguistic DataBase*

در محل‌های اتصال را کاهش می‌دهد. به طور خاصی، در مورد هجاهایی که با آواهای واکه‌ای آغاز شده و یا به پایان می‌رسند، مرز هجاها در محل عبور سیگنال از خط صفر گذشته و دوره‌های تناوب گام، به صورت کامل در محدوده‌ی هجا، قرار می‌گیرند. به عبارتی، مرز در ابتدای اولین و انتهای آخرین دوره‌ی تناوب گام، واقع شده است.

- در مورد هجاهایی که آوای انتهایی آن‌ها، واکه‌ای است، نقطه‌ای به عنوان پایان هجا در نظر گرفته شده که در آن زمان، در سیگنال ارتعاش تارآواها، حالت واک‌داری به پایان رسیده باشد.

در منبع [۱۴] از یک روش قدیمی ارائه شده در کتاب‌های انگلیسی زبان استفاده کرده است، به این صورت که مرزهای هجایی را به دو حالت کلی V.C و C.C در زبان فارسی تقسیم‌بندی کرده است. در هرکدام از آنها دامنه‌ی ارتعاش و ... ملاکی است برای جداسازی هجاها از یک‌دیگر. و با شرح و بسطی که درباره‌ی آواهای مختلف داده شده است تقسیم‌بندی‌ها را کوچکتر کرده است و برای هرکدام از آنها راه چاره‌ای داده است.

در [۲۰] با روش اتوماتیک و با استفاده از تبدیل موجک و ویژگی‌های مرز هجایی، عمل استخراج را انجام داده است. برای تقطیع هجایی سیگنال گفتار، ابتدا باید تابع انرژی زمان‌کوتاه، به روشی مناسب محاسبه شود. پس از آن، نوسانات ناخواسته به روش فازی، حذف شده و در پایان با استفاده از معیارهای آستانه، مرزهای هجایی آشکار می‌شوند. که در این مقاله، از روش آستانه‌گذاری ضرایب موجک، برای آشکارسازی مرزهای هجایی، استفاده شده است.

۳-۱۰-۲. استخراج دایفون

تهیه و طراحی پیکره‌ی زبانی لازم جهت استخراج دایفون‌ها در [۱۹]:

- ابتدا با مراجعه به واژگان، برنامه‌ی بازسازی گفتار فارسی (گویا)، برای هر یک از دایفون‌ها، سه کلمه را که دارای بیش‌ترین بسامد در واژگان برنامه، و شامل دایفون مورد نظر نیز باشد، استخراج می‌شود.
- برای برخی دایفون‌ها، کلمه‌ای پیدا نشد و از ترکیب کلمات و یا حتی کلمات بی‌معنی، استفاده شد.

در طی کار طراحی پیکره، به قواعد زیر توجه شود:

- تمام دایفون‌ها از هجاهای بی‌تکیه استخراج شده‌اند.
- هیچ یک از دایفون‌ها از داخل کلمات تک‌هجایی استخراج نشده‌اند.
- در فرآیند استخراج دایفون مورد نظر، واج اول و دوم دایفون، هر یک در دو هجای متفاوت نباشند. مثلا "شن" که یک دایفون "CC" است، از کلمه‌ی "پیشنهاد" جدا نمی‌شود چون این دایفون از دو هجای جداگانه باید استخراج شود ولی در کلمه‌ی "جشنواره" دایفون در یک هجا وجود دارد.
- برای استخراج خوشه‌های دوهمخوانی، سعی شده است که همخوان اول در داخل هجای "CVC" باشد و نه داخل هجای "CVCC".
- برای استخراج "CV" باید هجای قبل از آن نیز ترجیحا همان "CV" باشد نه "CVC".
- ...

اصول استخراج در این مقاله:

- دایفون‌هایی که با واکه شروع می‌شوند یا به واکه ختم می‌شوند، از وسط منطقه‌ی پایدار واکه برش زده شده است.
- دایفون‌هایی که با همخوان سایشی شروع یا به آن ختم می‌شوند، از وسط همخوان برش زده شده است.

- هم‌خوان‌هایی که با هم‌خوان انفجاری بی‌واک شروع می‌شوند، قسمت رهش (همراه با سایش) و دمش هم‌خوان آغازین دایفون را انتخاب کرده است.
- دایفون‌هایی که به هم‌خوان انفجاری ختم می‌شوند، قسمت بست (سکوت) هم‌خوان، انتخاب شده است.
- در تمام مراحل فوق، دایفون شامل قسمت گذر بین دو آوای سازنده‌ی خود است.
- برش سیگنال‌ها همواره از نقطه‌ی عبور از صفرشان صورت گرفته است.

در [۲۱] آمده است:

در این مقاله روش دستی را برای مرزگذاری دایفون‌ها پیشنهاد داده است. برای دقت بیشتر پیشنهاد شده است که از سه روش شنیداری، دیدن شکل موج زمانی، و استفاده از طیف نگاشت سیگنال (اسپکتروگرام) که اطلاعات زمان، فرکانس و انرژی را نشان می‌دهد.، به طور هم‌زمان استفاده شود.

چند نکته در هنگام تقطیع ذکر شده است. مثلا فایل‌های صوتی باید بر اساس نام دایفون نام‌گذاری شوند. یا این که در عمل مناسب‌تر است که هنگام جداسازی سیگنال مربوط به دایفون مورد نظر از کل سیگنال ضبط شده، بازه زمانی تقریبی ۵۰ میلی ثانیه قبل و بعد دایفون هم انتخاب شود.

در این پایان‌نامه نیز برای تقطیع از روش همین مقاله استفاده کرده‌ایم؛ ولی در جزئیات و ملاک‌های تقطیع دایفون اختلاف نظرهایی وجود دارد.

۳-۱۰-۳. روش استخراج SBS

در این مقاله، برای جداسازی واحد گفتار، از سکوت آن‌ها استفاده می‌شود. یعنی واحد جدیدی به نام SBS[1] معرفی کرده است که این دادگان بزرگ‌تر از هجا و کوچک‌تر از کلمه می‌باشند. مثلا کلمه‌ی مناظره ساختار هجایی چهارتکه دارد، ولی وقتی به نوع تلفظ این کلمه نگاه کنیم می‌شنویم: /منا/ و /ظره/، وقتی این دو تکه کنار هم قرار گیرند، گفتار بازسازی‌شده طبیعی‌تر خواهد بود نسبت به این‌که

[1] *Silence Based Segmentation*

چهار تکه‌ی هجایی کنار هم واقع شوند. در این روش واکه‌ها به دو صورت بلند "W" و کوتاه "V" بیـان شده‌اند. در یک متن هرجا که به مدل"CCW"، "CCV"، "WVW"، "WCV"برخورد شد، مکان تقطیـع همان‌جاست. مثلا در کلمه‌ی منتظر که مدلش به صورت "CVCCVCVC" می‌باشد، به شـکل "CVC" و "CVCV" تقطیع می‌شود. [۱]

فقط حدود ۴۰۰۰ هجا در فارسی وجود دارند که کارایی دارند. ولی با این تقسیم بنـدی بـه دادگـان چند برابری نسبت به این تعداد نیاز داریم. در روش SBS تعداد دادگـان زیـاد مـی‌شـود ولـی پـردازش کمتری برای اتصال لازم است و خروجی، طبیعی‌تر خواهد بود.

فصل چهارم:

روش پیشنهادی برای طبیعی‌سازی گفتار با استفاده از دادگان دایفونی هفت‌گانه

۴-۱. مقدمه

در این فصل، درباره‌ی کار عملی انجام شده در پایان‌نامه توضیح داده‌ایم. قدم اول در این کار، انتخاب واحد گفتاری لازم برای سنتز است. در فصل دوم توضیح مفصلی درباره‌ی واحدهای آوایی دادیم و دایفون را به عنوان واحد گفتار، مناسب دیدیم. مزیت دایفون، این است که تعداد آن نسبت به هجا خیلی کمتر است و چون دایفون‌ها به خوبی گذرهای آوایی را پوشش می‌دهند، در سنتز آن‌ها خروجی مناسبی ایجاد می‌شود. از نظر وضوح و قابل فهم بودن شاید به دادگان هجایی نرسند، ولی از نظر طبیعی بودن گفتار خروجی بسیار بهتر است، چون در سنتز به وسیله دادگان هجایی، دچار کوبه‌ای شدن گفتار بازسازی‌شده می‌شویم و طبیعی کردن آن نیازمند پردازش سیگنال پیچیده و سنگینی است [۱۹].

در فصل سوم، درباره‌ی انواع روش‌های استخراج واحدهای آوایی توضیح دادیم. برای استخراج دایفون از روش استفاده شده در [۲۱] که روش موازی شنیداری، دیدن شکل موج و توجه به طیف-نگاشت (اسپکتروگرام) سیگنال گفتار است، بهره می‌بریم. البته نگارنده‌ی پایان‌نامه، نکاتی را هنگام تقطیع رعایت کرده است که گاهی مخالف نظرات نویسنده‌ی این مقاله می‌باشد.

در ادامه این فصل با ذکر مثال و توضیح کافی می‌خواهیم خواننده را مجاب کنیم که می‌توان حالت‌های مختلف گفتاری را با استفاده از هفت حالت مختلف دایفونی، مدل کرد. برای این کار، انواع آهنگ جمله و تکیه کلمه با استفاده از علم نگارنده، تقسیم‌بندی می‌شود و یک حالت دایفونی نیز به فاصله‌ی میان‌واژه‌ای تعلق می‌یابد.

در نهایت نکات مهمی درباره شرایط و قوانین ضبط و استخراج دایفون‌ها بیان شده است و سپس با دید کلی‌تر این روش را برای تهیه دادگان دایفونی پیشنهاد می‌دهیم. برای استفاده از روش پیشنهادی برای تهیه‌ی دادگان دایفونی، باید مسائلی از قبیل هنجارسازی سنتز، نحوه‌ی نام‌گذاری و جستجوی دادگان و ... مورد نظر قرار بگیرند که برخی از آن‌ها را نام برده و توضیح مختصری درباره حل آن‌ها بیان کرده‌ایم.

۴-۲. مزایای دایفون و دلایل استفاده از آن در روش پیشنهادی

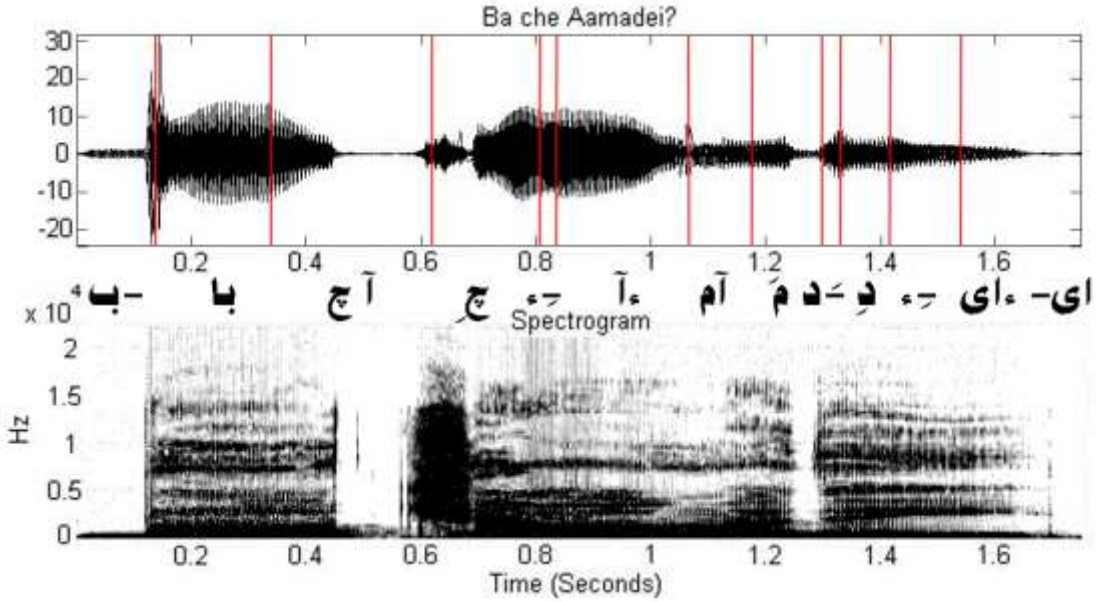
تحقیقات نگارنده در اجزای سخن در زبان فارسی، (که بخشی از آن‌ها در فصل سوم آورده شده است) به این نتیجه رسید که دایفون‌ها واحدهای کوچک و در عین حال دارای خاصیت‌های زیادی هستند که می‌توان آن را بهترین واحد زبان برای سنتز دانست. در این قسمت مزایا و دلایل استفاده نگارنده از این واحد گفتار را بیان می کنیم.

- پیوستگی: این نمونه‌ها از گفتار پیوسته استخراج می‌شوند و به همین دلیل در سنتز این تکه-های صوتی نیز پیوستگی خواهیم داشت.
- تعداد کم دایفون‌ها: تعداد کم آن‌ها باعث می‌شود که زمان زیادی برای استخراج آنها نسبت به واحدهای دیگری مانند هجا، ترایفون و کلمه لازم نباشد. هم‌چنین حجم دادگان آن نیز نسبتا کمتر خواهد بود و به همین دلیل سرعت پاسخ‌دهی سیستم نیز بالا خواهد رفت.
- صدای انسان: دایفون‌ها از صدای طبیعی انسان استخراج می‌شوند. البته بعضی روش‌ها هستند (در فصل دوم به آنها اشاره شده است) که از صدای مصنوعی تولید شده از کامپیوتر استفاده می‌کنند که باعث ناخوشایند شدن گفتار سنتز شده می‌شوند.
- مدل‌سازی تکیه: تکیه مربوط به شدت‌دار کردن یک یا چند هجا در کلمه است و از آنجایی که رابطه بین دایفون و هجا یک رابطه واضح است، می‌توان به سادگی تکیه‌دار بودن یا نبودن یک هجا را با استفاده از دادگان دایفونی، مدل سازی کرد.
- مدل‌سازی آهنگ: در این مورد نیز، با توجه به این که رابطه‌ی بین دایفون و هجا و کلمه و جمله مشخص است به راحتی می‌توان توسط مدل‌های دایفونی، جمله‌ی با آهنگ دلخواه را تولید کرد.

۴-۳. روش استخراج دایفون‌ها

برای استخراج دایفون ابتدا به یک پیکره نیاز داریم. این پیکره باید به گونه‌ای باشد که تمام دایفون-های مورد نیاز برای استخراج را دارا باشد. این پیکره باید شامل هفت مدل دایفونی (که در همین فصل راجع به آن توضیح خواهیم داد) باشد. بهتر است پیکره به صورت مجموعه‌ای از جملات باشد.

برای تقطیع دایفون‌ها، در این پایان‌نامه از سه روش به صورت موازی استفاده شده است که می-توانند کار استخراج دایفون را راحت‌تر و دقیق‌تر کنند. در ابتدا از روش شنیداری برای جداسازی اولیه استفاده می‌شود. سپس با توجه به شکل موج سیگنال و اسپکتروگرام[1] آن، تقطیع بهبود می‌یابد. شکل زیر، تقطیع یک جمله به دایفون‌های تشکیل دهنده‌اش را نشان می‌دهد که البته شاید همه‌ی آن‌ها، برای ذخیره در دادگان دایفونی مناسب نباشند؛ چرا که دایفون‌ها باید دارای شرایط خاصی باشند که بتوان آن‌ها را در دادگان ذخیره کرد.



شکل ۴-۱: بهبود تقطیع دایفون‌ها به وسیله‌ی اسپکتروگرام

[1] اسپکتروگرام یا طیف‌نگاشت‌هایی که در این پایان نامه آورده‌ایم به این صورت است که مناطقی که سفید هستند دارای کمترین انرژی و هرچه به سمت رنگ مشکی نزدیک می‌شویم انرژی بیشتری دارند. محور افقی زمان و محور عمودی فرکانس می‌باشد.

در شکل ۴-۱ اسپکتروگرام، دیده می‌شود که هر کدام از واج‌ها دارای فرکانس و انرژی خاصی هستد. مثلا حروف صدادار، فرکانس ثابت و پایینی دارند و دارای انرژی بیشتری هستند ولی حروف بی‌صدا، دارای فرکانس بالا و انرژی کمتری نسبت به حروف صدادار می‌باشند. با دانستن خواص حروف می‌توان آن‌ها را در نمودار اسپکتروگرام بهتر درک نمود. مثلا حروف خشیومی (Nasal) مانند حرف "م" که به مسدود شدن دهان مربوط هستند در نمودار اسپکتروگرام خود، دارای منطقه‌ی صفر می‌باشند و کاملا در نمودار، حالت بارزی دارند. همانطوری که در شکل بالا دیده می‌شود حرف "م" دارای قسمتی است که ناحیه‌ی صفر می‌باشد و در آن فرکانس‌ها مقدار خیلی کمی دارند.

در بخش بعد به قواعد و مقرراتی اشاره می‌شود که در استخراج دایفون‌ها باید مد نظر داشت تا بتوانیم دادگان مفیدی برای سیستم سنتز آماده کنیم.

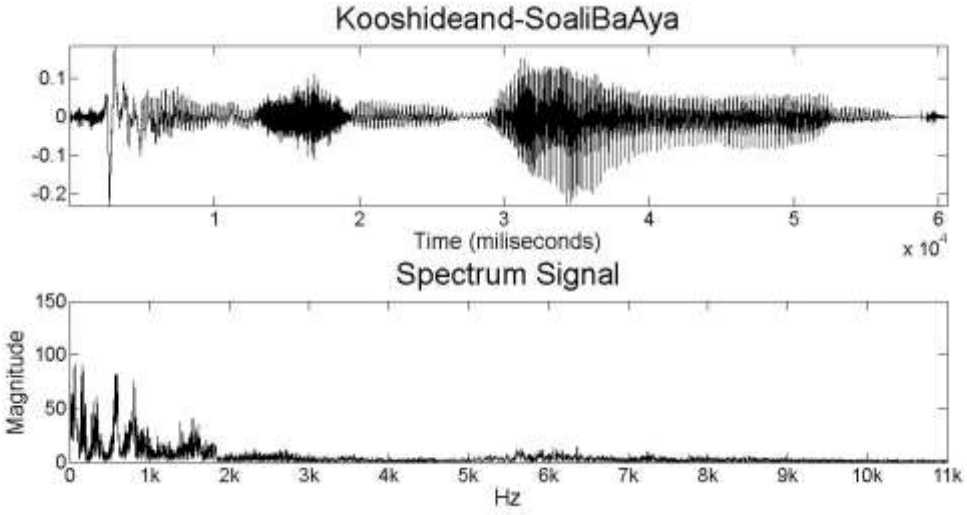
۴-۴. آهنگ جملات

با توجه به آنچه در مبانی زبان‌شناسی فصل سوم گفته شد، می‌خواهیم جمله‌های زبان فارسی را از نظر آهنگ گفتار به دسته‌های کلی‌تری تقسیم کنیم. در این پایان‌نامه سعی شده است که آهنگ جمله را فقط روی فعل آن‌ها اعمال کنیم و از بقیه‌ی اقسام جمله صرف نظر کنیم؛ چون هر کدام از اقسام جمله، می‌توانند آهنگ خاصی بگیرند و از آن‌جا که حجم دادگان و زحمت تهیه‌ی آن‌ها خیلی زیاد می‌شود، از تهیه‌ی دادگان برای اقسام دیگر جمله صرف نظر شده است. البته تقسیم‌بندی آهنگ قسمت‌های دیگر جمله، کاری بسیار دشوار و تقریبا محال است.

انواع فعل‌های فارسی را از لحاظ آهنگ گفتار می‌توان به چند دسته‌ی کلی و پرکاربرد تقسیم‌بندی کرد و با فعل "کوشیده‌اند" که در جملات خوانده شده جهت استخراج دایفون وجود دارد، مقایسه‌ی بین این حالات انجام گرفته است:

۱- جمله‌ی سوالی با "آیا": جمله‌هایی که پاسخ آن‌ها "بله" یا "خیر" خواهد بود. با تحقیقاتی که بر روی این افعال انجام دادیم، به این نتیجه رسیدیم که آهنگ این جملات "خیزان" است،

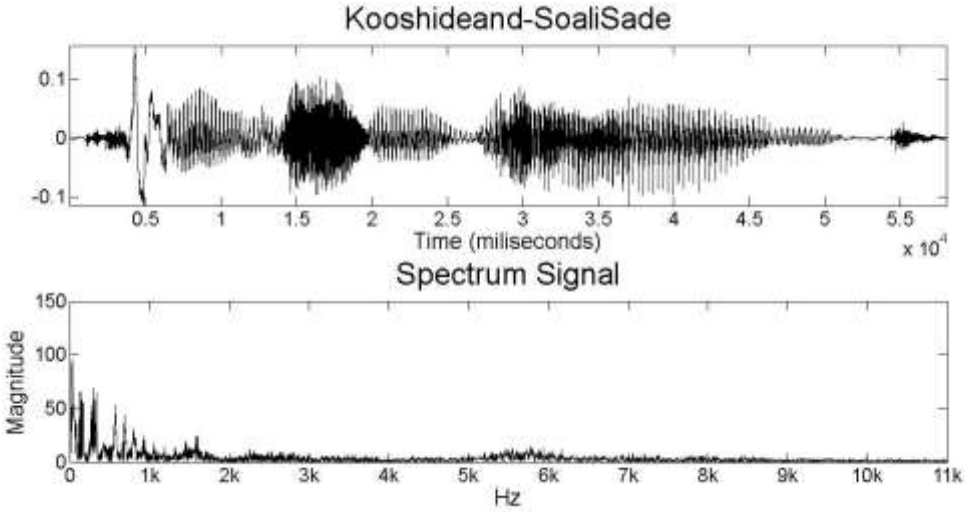
یعنی از صدای پایین و فرکانس گام پایین شروع می‌شوند و هر چه به انتهای جمله نزدیک می‌شویم این مقادیر بیشتر می‌شوند و آخرین هجای این افعال، دارای بیش‌ترین فرکانس و انرژی می‌باشند. با توجه به شکل ۴-۲ می‌توانیم نظریه‌ی خود را بیشتر اثبات کنیم.



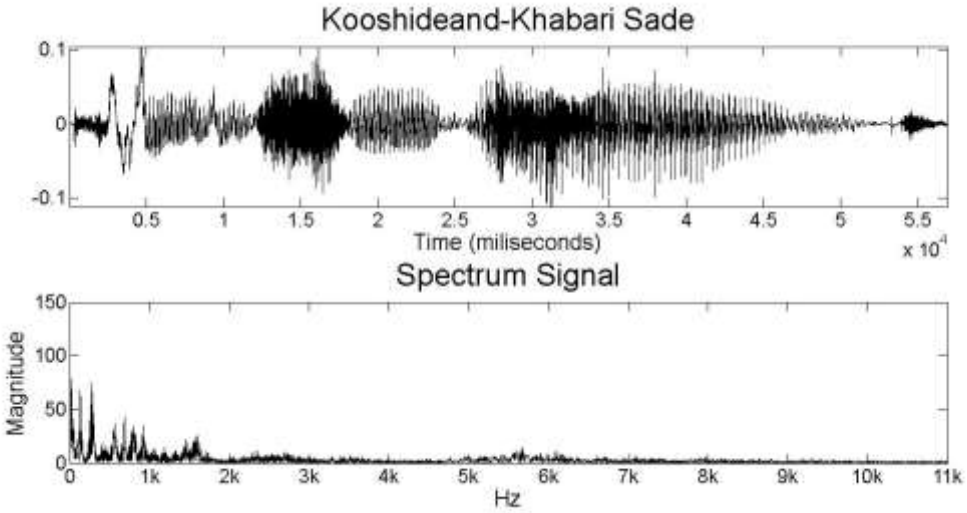
شکل ۴-۲: فعل "کوشیده‌اند؟" با واژه‌ی پرسشی "آیا" شکل موج سیگنال و تبدیل فوریه‌ی سیگنال (فرکانس‌های سازنده‌ی سیگنال)

با توجه به شکل ۴-۲ در فعل سوالی با واژه‌ی "آیا" معمولا فرکانس‌های بالای بیشتری نسبت به انواع فعل‌های دیگر وجود دارد. البته فعل سوالی-تعجبی دارای فرکانس‌های بیشتری نسبت به تمام انواع فعل است.

۲- جمله‌ی سوالی و خبری: جملات سوالی می‌توانند با کلمات پرسشی دیگری نیز آغاز شوند که اکثرا با "چه" آغاز می‌شوند. بعضی جملات پرسشی نیز شاید، دارای واژه‌ی پرسش نباشند. جملات خبری ساده از نظر بیانی، بسیار شبیه به جملات پرسشی هستند؛ از این رو در تقسیم‌بندی حالات مختلف جمله، فعل آن‌ها را در یک گونه‌ی آهنگی یکسان قرار می‌دهیم. در این گونه افعال، بیش‌ترین تاثیر روی آهنگ جمله را، هجای آغازین فعل بر عهده دارد.



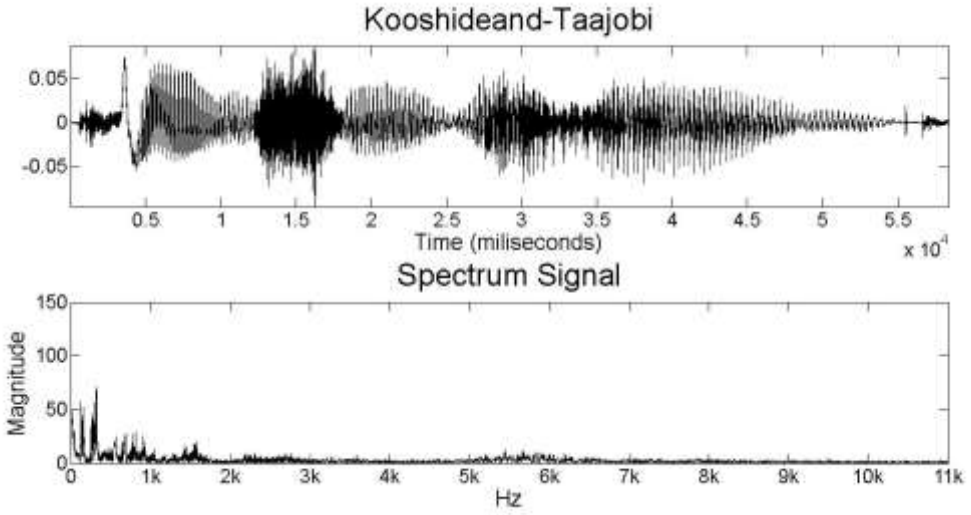
شکل ۴-۳: فعل سوالی ساده "کوشیده‌اند؟" شکل موج سیگنال و تبدیل فوریه‌ی سیگنال (فرکانس‌های سازنده‌ی سیگنال)



شکل ۴-۴: فعل خبری ساده "کوشیده‌اند." شکل موج سیگنال و تبدیل فوریه‌ی سیگنال (فرکانس‌های سازنده‌ی سیگنال)

با توجه به آن‌چه در شکل‌های ۴-۳ و ۴-۴ دیده می‌شود به این نتیجه می‌توان رسید کـه شـکل موج و فرکانس‌های دربردارنده‌ی این افعال، بسیار شبیه به هم هستند. پس این افعال را با هـم در یک گروه واحد قرار می‌دهیم.

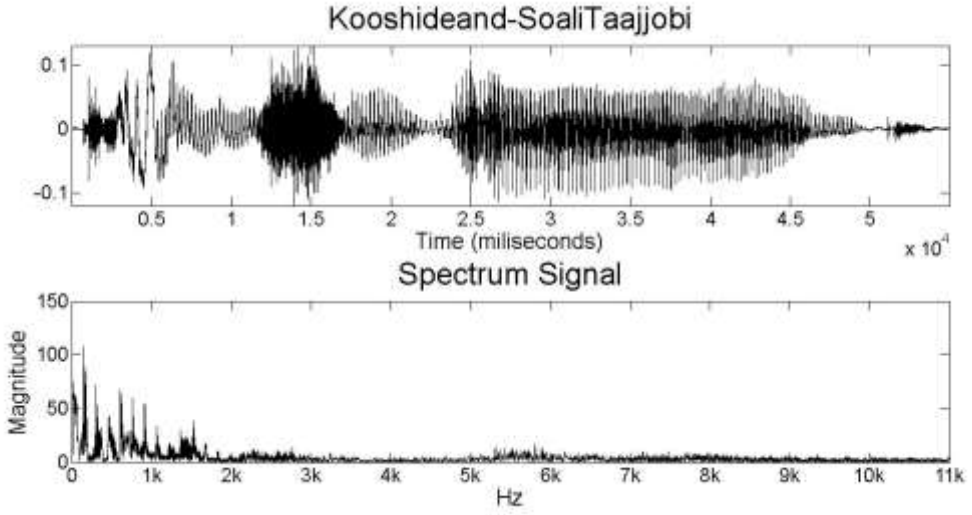
۳- جمله‌ی تعجبی: در این گونه جملات، بیش‌ترین تاثیر آهنگین جمله بر روی هجـای اول فعـل آن می‌باشد. فعل جمله‌ی تعجبی فرکانس پایین دارد و سهم فرکانس‌های بـالا در آن، نسـبت به انواع دیگر کمتر است.



شکل ۴-۵: فعل تعجبی "کوشیده‌اند!" شکل موج سیگنال و تبدیل فوریه‌ی سیگنال (فرکانس‌های سازنده‌ی سیگنال)

این افعال دارای انرژی و فرکانس پایین هستند. همانطور که در شـکل ۴-۵ مشـاهده مـی‌شـود انرژی سیگنال شکل مـوج آن بسـیار کمتـر از بقیـه‌ی افعـال اسـت و همچنـین ایـن فعـل دارای فرکانس‌های پایین است و فرکانس‌های بالا در تبدیل فوریه‌ی آن کمتر به چشم می‌خورد.

۴- جمله‌ی سوالی-تعجبی: این گونه جملات باید دارای لحنی باشند که بـه طـور هـم‌زمـان، هـم مفهوم تعجب را بفهمانند و هم سوال در آن‌ها باشد. از ایـن رو طبـق تحقیقـاتی کـه در زبـان فارسی انجام شده است، نگارنده به این نتیجه رسیده است که هجـای آغـازین ایـن افعـال بـا حالت تعجبی و هجای پایانی آن‌ها با حالت سوالی همراه است. البته منظور از سوالی در ایـن-جا، سوالی با واژه‌ی "آیا" می‌باشد. ولی باز هم از آنجایی که این افعال حالـت ثـابتی ندارنـد و افراد مختلف با لحن‌های مختلف آن‌ها را بیان می‌کنند نمی‌توان بـه طـور قطعـی بـرای آن‌هـا تصمیم‌گیری کرد.



شکل ۴-۶: فعل سوالی-تعجبی "کوشیده‌اند؟!" شکل موج سیگنال و تبدیل فوریه‌ی سیگنال (فرکانس‌های سازنده‌ی سیگنال)

همانطور که در شکل ۴-۶ مشهود است این فعل فرکانس‌های بالای بیشتری نسبت به افعال دیگر دارد. با یک بررسی اجمالی روی شکل موج افعال تعجبی و سوالی با "آیا" به این نتیجه دست پیدا می‌کنیم که در این فعل، هجای نخست شبیه به هجای ابتدایی فعل تعجبی است و هجای پایانی آن شبیه به هجای آخر فعل سوالی با واژه‌ی "آیا" است.

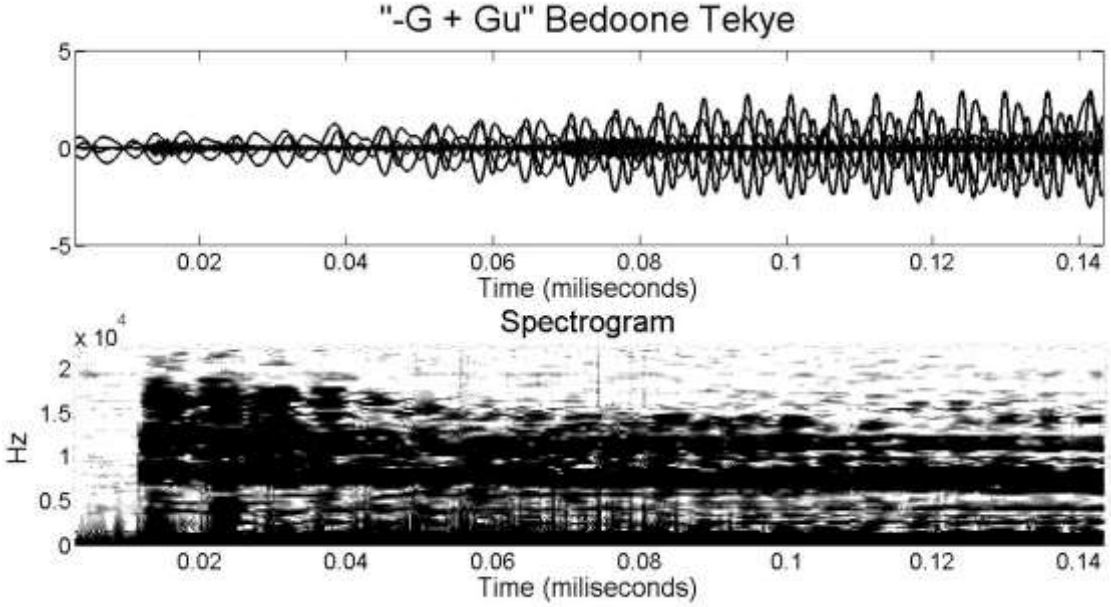
۴-۵. تکیه‌ی کلمات در جمله

با توجه به آن‌چه در بحث زبان‌شناسی آمده است، تکیه‌ی کلمات در جمله، ممکن است تغییر کند و یا کم و زیاد شود و یا حتی تکیه‌ی بعضی کلمات حذف خواهد شد. پس در قسمت پردازش متن سیستم متن به گفتار، باید به این نکات توجه شود و پایگاه متنی وجود داشته باشد که به وسیله‌ی آن، تکیه‌ی کلمات در جمله مشخص شود. در این پایان‌نامه با فرض این که در کلمات، مکان تکیه مشخص شده است، در قسمت سنتز گفتار، دایفون‌های تکیه‌دار یا بدون تکیه را در جایگاه مناسب خودشان قرار

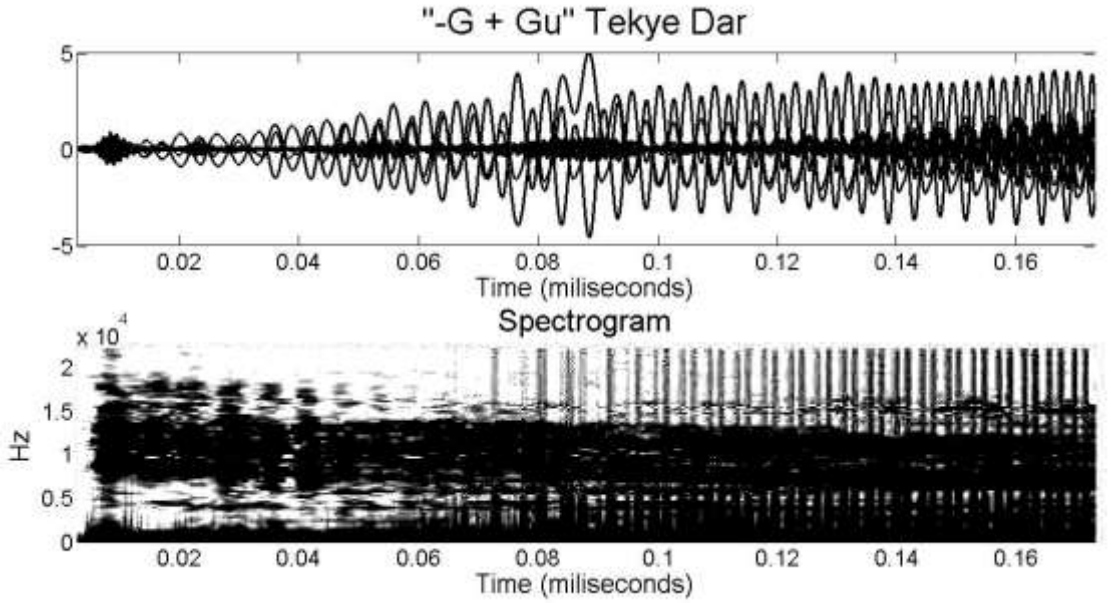
می‌دهیم. در خروجی، ملاحظه می‌شود که هجای تکیه‌دار به زیبایی از هجاهای بدون تکیه، تمییز داده می‌شوند و اگر این‌گونه نبود، گفتار سنتز شده، یکنواخت و غیر طبیعی به گوش می‌رسید.

در شکل‌های ۴-۷ و ۴-۸ سیگنال موج و اسپکتروگرام هجای "گو" در واژه‌ی "گویا" نشان داده شده است. این واژه در دو معنی مختلف، در جمله‌ی "گویا، مطالب گویا است." بیان شده است. در شکل ۴-۷ هجای "گو" بدون تکیه می‌باشد و در شکل ۴-۸ هجای "گو" تکیه‌دار می‌باشد. می‌توان به راحتی تشخیص داد که این هجا در حالت تکیه‌دار دارای انرژی و فرکانس‌های بالای بیشتری می‌باشد و همچنین طول مدت بیان این هجا، نسبت به حالت بدون تکیه بیشتر است. همانطور که در [۱۸] درباره‌ی تکیه و ویژگی‌های آن توضیح داده شده، طول زمان هجای تکیه دار ۱/۳ برابر طول زمان هجای بدون تکیه است. همچنین نکات دیگری که در آن‌جا بیان شده‌اند را نیز می‌توان کاملا مشاهده و درک نمود.

در روش پیشنهادی، به جای مدل‌های پیچیده‌ی ریاضی، برای مدل کردن هجای تکیه‌دار، از گفتار طبیعی و تقطیع آن استفاده می‌شود. به این صورت که هجای مورد نظر را به صورت پیوسته و با حالت‌های با تکیه و بدون تکیه ضبط می‌کنیم و سپس دایفون‌های تشکیل دهنده‌اش را استخراج و در دادگان ذخیره می‌کنیم. با این کار می‌توانیم در هنگام سنتز، بدون هیچ عملیات پیچیده‌ی ریاضی، گفتار طبیعی‌تری در خروجی داشته باشیم.



شکل ۴-۷: سیگنال موج و اسپکتروگرام هجای "گو" در کلمه‌ی "گویا" (تکیه روی هجای "یا")



شکل ۴-۸: سیگنال موج و اسپکتروگرام هجای "گو" در کلمه "گویا" (تکیه روی هجای "گو")

۴-۶. دایفون‌های میان واژه‌ای

برای پیوستگی سخن در این پایان‌نامه پیشنهاد شده است، که دایفون‌هایی که بین کلمات یک جمله وجود دارند را استخراج کنیم و هنگام سنتز از آن‌ها استفاده کنیم. می‌توان به راحتی دریافت که وجود این دایفون‌ها چقدر در طبیعی‌تر و روان‌تر شدن جملات سنتز شده نقش موثری ایفا خواهند کرد.

مقایسه‌ی وجود این دایفون‌ها و عدم وجود آن‌ها را در فصل بعد انجام داده‌ایم و از نظر پیوستگی شکل موج و طبیعی بودن گفتار خروجی آن‌ها را مورد بررسی قرار داده‌ایم.

۴-۷. انواع دایفون‌های مورد نیاز برای پروژه

با توجه به آن‌چه در بخش قبلی گفته شد، دایفون‌های مورد استخراج به هفت حالت تقسیم می‌شوند که باید برای هر کدام از حالت‌ها، تمام دایفون‌ها را استخراج کنیم. این حالت‌ها شامل موارد زیر می‌باشد:

- دایفون‌های لازم برای هجای ابتدایی فعل سوالی با واژه‌ی "آیا"
- دایفون‌های لازم برای هجای پایانی فعل سوالی با واژه‌ی "آیا"
- دایفون‌های لازم برای هجای ابتدایی فعل خبری و پرسشی
- دایفون‌های لازم برای هجای ابتدایی فعل تعجبی
- دایفون‌های لازم برای هجای تکیه‌دار
- دایفون‌های لازم برای هجای بدون تکیه
- دایفون‌های میان واژه‌ای

از آن‌جا که روش پیشنهادی وقت‌گیر است و از طرفی این کار نیاز به استودیوی ضبط صدا و امکانات جانبی زیادی دارد، در این پایان‌نامه به ذکر مثال‌هایی برای اثبات فرضیه‌ی خود، بسنده کرده‌ایم. ابتدا قوانینی باید ذکر شود که بتوان طبق آن‌ها متن مورد نیاز را نوشت و آن را ضبط کرد، البته این قوانین، باید بیش‌تر و مطمئن‌تر شوند که نمونه‌های استخراج شده در بخش سنتز، خوب عمل کنند؛ ولی با توجه به علم اندک زبان‌شناسی نگارنده، قوانین ساده و کارآمدی وضع شده و استخراج صورت پذیرفته است. همچنین در فصل دوم، قواعدی درباره‌ی ساختار متن و استخراج دایفون از چند مقاله آمده است که از آن‌ها استفاده کردیم. ولی باز هم باید در نظر داشت که این نظریات کافی

نیست و نیاز به علم و تجربه‌ی بیشتری دارد که بتوانیم در قسمت سنتز با یک خروجی خوب و مناسب مواجه شویم.

متن حاضر شده برای ضبط را به گونه‌ای انتخاب کردیم که دارای تمام دایفون‌های مورد نیاز برای قسمت سنتز باشد. در قسمت سنتز، چهار جمله با حالت‌های مختلف سوالی با واژه‌ی "آیا"، خبری، تعجبی، سوالی-تعجبی را می‌بایست تولید کنیم. نتیجه‌ی این آزمایشات را در فصل بعد آورده‌ایم و توضیحات لازم را در مورد آن‌ها کامل کرده‌ایم.

۴-۸. قوانین حاکم بر استخراج دایفون‌ها

۴-۸-۱. محل تقطیع آواها

در یک ترکیب cvc از جایی که از نظر زمانی تقریبا در وسط حرف صدادار است تقطیع را انجام می‌دهیم و معمولا محل تقطیع جایی قرار می‌گیرد که بیشینه حرف صدادار است. به همین صورت زمانی که vcv است و می‌خواهیم تقطیع را انجام دهیم از کمینه‌ی حرف بی‌صدا برای این کار استفاده می‌کنیم و سعی می‌کنیم وسط بودن محل تقطیع را تا حد زیادی رعایت کنیم. البته باید این نکته را در نظر داشت که باید محل تقطیع، جایی تقریبا در وسط طول سیگنال آوای مورد نظر باشد.

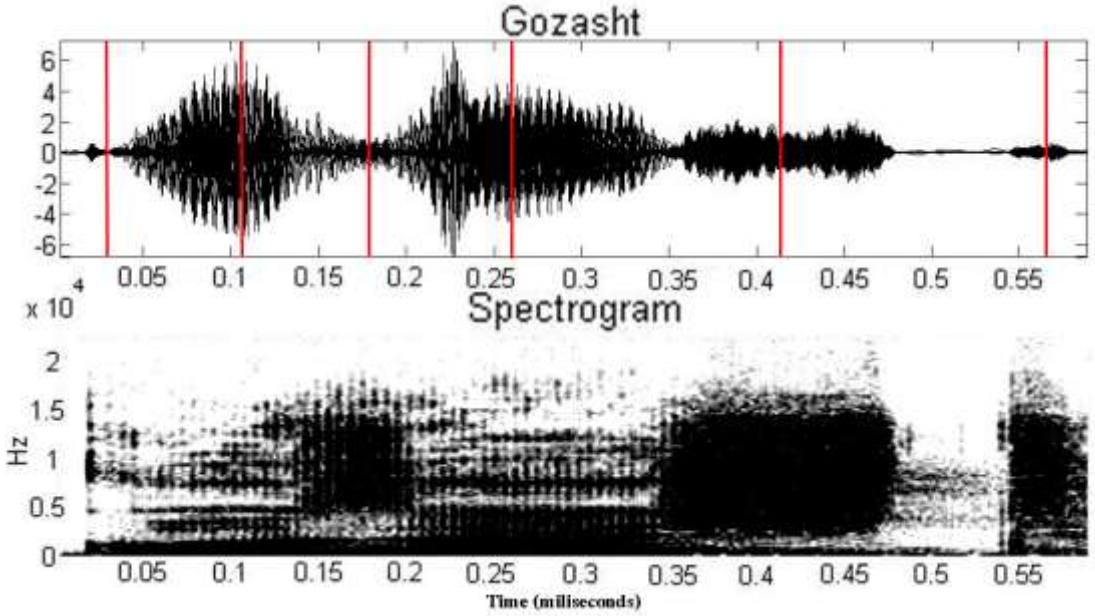
البته قوانین دقیق‌تری نیز می‌توان رعایت کرد (مثل قوانینی که در [۱۹] رعایت کرده است) که دادگان بهتری داشته باشیم ولی از آن‌جا که این اطلاعات، بیشتر به علم زبان‌شناسی وابسته است، از آن‌ها صرف نظر کردیم.

شرایط بیشتری برای تقطیع رعایت کرده‌ایم، که چند مورد را به اختصار بیان می‌کنیم:

- در تمام انواع دایفون‌ها و شیوه‌ی استخراج آن‌ها یک نکته‌ی کلی باید رعایت شود و آن، این است که دایفون شامل قسمت گذر بین دو آوای سازنده‌ی خود باشد.
- در تقطیع باید رعایت شود که همواره از نقطه‌ی عبور از صفرشان، تقطیع صورت گیرد.

- در بعضی مقالات از قسمت تقطیع تا جایی که می‌بایست قطع شود فاصله‌ی کوتاهی است. در اصل باید، هم‌پوشانی دو قطعه‌ی جدا شده صورت گیرد ولی در این پایان‌نامه این روش توصیه نمی‌شود چون در هنگام سنتز دچار طولانی شدن دایفون‌ها و غیر طبیعی شدن خروجی می‌شویم.

شکل ۴-۹ از این قوانین در استخراج دایفون‌ها استفاده کرده است و رعایت بیشتر آن‌ها را می‌توان به وضوح در آن دید.



شکل ۴-۹: تقطیع دایفون‌های واژه‌ی "گذشت" با استفاده از اسپکتروگرام و کمینه و بیشینه‌ی آواها در شکل موج

۴-۸-۲. ساختار دایفونی جملات

هجاها باید دارای ساختار دایفونی خاصی باشند. مثال آن در زیر آمده است:

- ساختار جمله ای:

چقدر زود آمده‌ای!

- هجاهای جمله:

چِ قدر زود آ مَ دِ ای

- دایفون‌های جمله همراه با تقسیم بندی هجایی:

-چ چِ (دایفون‌های بدون تاکید) ِق قَ َد در (دایفون‌های تاکیـددار) ر ز (دایفـون-های میان‌واژه‌ای) زو اود (دایفون‌های تاکیددار) د ء (دایفون‌های میان‌واژه‌ای)

ءآ (هجای ابتدایی جمله‌ی تعجبی) آم مَ (دایفون‌های بـدون تاکیـد) َد دِ (دایفـون-های بدون تاکید) ِء ءای (دایفون‌های بدون تاکید)

۴-۸-۳. نام‌گذاری دادگان دایفونی

برای نام‌گذاری دادگان می‌توان از قواعد و قوانینی که قبلا وضع شده مثل "الفبـای آوانگـار بـین‌الملـل IPA" یا روش "سـمپا X-SAMPA" اسـتفاده کـرد و یـا مثـل دادگـان گفتـاری "فـارس‌دات"، الفبـای آوانویسی خاص خودمان را داشته باشیم. البته در این تحقیق، چون دادگان مورد نیاز ما، زیاد نبودند از حروف انگلیسی مناسب برای ذخیره‌ی دادگان استفاده کردیم.

۴-۹. شرایط ضبط صدا

نکاتی در مورد ضبط صدا وجود دارد که علم به آنها ما را در پیشرفت مطالعاتمان بسیار کمک می‌کنـد. که به اختصار چند مورد از آن‌ها را توضیح می‌دهیم:

۴-۹-۱. دستگاه ضبط مورد استفاده

ضبط صدا باید در محیط مناسب و توسط دستگاه‌های خوب با کمتـرین نـویز باشـد. بـرای ایـن کـار، دستگاه ضبطی تهیه شد که دارای مشخصات بسیار عالی برای این گونه فعالیت‌های تحقیقی می‌باشـد. این دستگاه، ساخت شرکت SONY با مدل PCM-M10 می‌باشد که در شکل ۴-۱۰ آمده است.

شکل ۴-۱۰: دستگاه ضبط حرفه‌ای Sony PCM-M10

۴-۹-۲. حالت گوینده

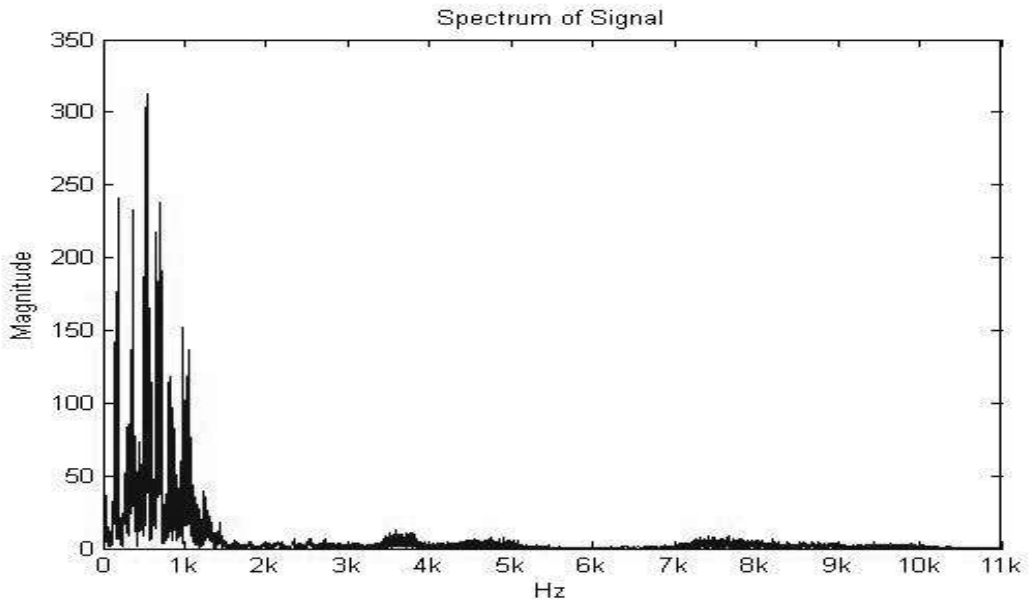
گوینده‌ی متن باید فاصله‌ی دهان تا میکروفون را ۲۰-۲۵ سانتی متر ثابت نگه‌دارد. سعی شود متن در دفعات کمی ضبط شود که حالت و نوع صدای گوینده یکسان باشد. فرد گوینده باید سعی کند تا در مواردی مانند فاصله‌ی میان‌هجایی و میان‌کلمه‌ای، سرعت، شدت، حالت بیان و ... از الگوی ثابتی استفاده کند. این الگوها مربوط به خود گوینده است و باید سعی شود در طول ضبط این حالت را حفظ کند. فرد گوینده باید لهجه‌ی فارسی تهرانی داشته باشد و متن را روان و بدون مکث اضافی و با رعایت تکیه‌ها و حالت‌های جمله بیان کند.

موارد دیگری در بیان نوع ضبط و فرد گوینده وجود دارد که تمام آن‌ها لازم به ذکر نیست و البته بسیاری از آن‌ها را نگارنده‌ی پایان‌نامه نیز نتوانسته است رعایت کند.

۴-۹-۳. سرعت نمونه‌برداری ضبط گفتار

سرعت نمونه‌برداری نه تنها روی حجم اشغالی تاثیر می‌گذارد، بلکه به پردازنده‌ای با مشخصات فرکانسی مورد نظر، نیاز دارد. طبق نرخ نایکوئیست برای نمونه‌برداری از سیگنال پیوسته، به فرکانسی

بزرگتر یا مساوی دوبرابر بیش‌ترین فرکانس موجود در سیگنال اولیه نیاز است. بیشترین فرکانس موجود در صدای انسان ۲۰ کیلوهرتز است. پس بیشترین کیفیت صوتی که برای ضبط گفتار نیاز است فرکانس ۴۰ کیلوهرتز می‌باشد، که در این پروژه داده‌ها را با فرکانس ۴۴۱۰۰ هرتز ضبط کرده-ایم. البته در بسیاری از کاربردهای مخابراتی، حتی فرکانس ۴ کیلوهرتز نیز کفایت دارد. ولی از آنجا که از نظر آماری بیش از ۹۶ درصد محتویات فرکانسی سیگنال‌های صوتی در پهنای باند حدود ۱۱ کیلوهرتز قرار دارد، فرکانس نمونه‌برداری ۲۲ کیلوهرتز برای یک پروژه‌ی کلی و بزرگ، بسیار عالی است. شکل ۴-۱۱ نشان می‌دهد که بیشترین فرکانس‌های موجود در یک تکه گفتار در محدوده ۰ تا ۱۱ کیلوهرتز می‌باشد.



شکل ۴-۱۱: فرکانس‌های موجود در سیگنال گفتار سه ثانیه‌ای

۴-۹-۴. فشرده‌سازی نمونه‌ها

فشرده‌سازی باعث می‌شود حجم اشغالی کمتر شود ولی میزان پردازش افزایش می‌یابد. پس با توجه به این که میزان حافظه‌ی مجاز اشغالی توسط دایفون‌ها، سرعت پردازنده، استفاده‌ی آنلاین یا آفلاین چگونه باشد، درباره‌ی این قضیه تصمیم گرفته می‌شود. باید این نکته را در نظر داشت که بیشتر

اوقات، در یک فشرده‌سازی کیفیت سیگنال پایین می‌آید و فشرده‌سازی بیشتر سیگنال باعث کاهش بیشتر کیفیت سیگنال می‌شود.

در این پروژه با توجه به این‌که سرعت پردازنده‌های کنونی بسیار بالاست و دسترسی به مقادیر بالایی از حجم حافظه امکان‌پذیر است، از دادگان با فرمت wav. استفاده کردیم.

۴-۹-۵. کاهش نویز

حضور نویز، در داده‌های صوتی که دامنه‌ی کمتری دارند، نمود بیشتری دارد. مثل دایفون "–س" که دارای دامنه‌ی کمی می‌باشد و وجود نویز با دامنه‌ی بالا در این دایفون باعث اختلال و خرابی آن می‌شود. استفاده از اتاق‌های آکوستیک برای ضبط، برای کاهش نویز بسیار مفید می‌باشد.

۴-۹-۶. عمق حافظه

پارامتر دیگری که در کیفیت سیگنال ضبط شده موثر است، عمق حافظه می‌باشد. مثلا اگر ضبط با عمق ۱۶ بیتی صورت بگیرد کیفیت بیشتری نسبت به عمق ۸ بیتی دارد، ولی این اختلاف کیفیت، توسط گوش انسان قابل تفکیک نیست.

بهتر است در پروژه‌ی انجام شده، فرکانس و عمقی داشته باشیم که برای گفتار مناسب و کافی باشد.

۴-۱۰. مسائل مهم در عملی کردن روش پیشنهادی در سیستم متن به گفتار فارسی

۴-۱۰-۱. ترتیب دایفون‌ها

در پردازش‌های آنلاین، سرعت عملکرد و دسترسی به داده‌ها بسیار اهمیت دارد. روشی که نگارنده در این پایان‌نامه برای ترتیب جستجو در دایفون‌ها پیشنهاد می‌دهد به این صورت می‌باشد که طبق احتمال قرار گرفتن دایفون‌ها در کلمات اولویت‌بندی می‌کند. مثلا اولین دایفونی که در کلمه قرار می-

گیرد C– است و آخرین دایفون به صورت -C یا -V خواهد بود که حالت -C محتمل‌تر است. بـه همـین ترتیب دومین دایفون همیشه CV است. به طور کلی اولویت قرار گرفتن دایفون‌ها به شکل زیر خواهـد بود:

۱- -C
۲- CV
۳- VC
۴- CC
۵- C-
۶- V-

اکنون باید نوع ترتیب قرار گرفتن صامت‌ها و مصوت‌ها را مشخص کرد. دو حالت را مـی‌تـوان بـرای این کار در نظر گرفت:

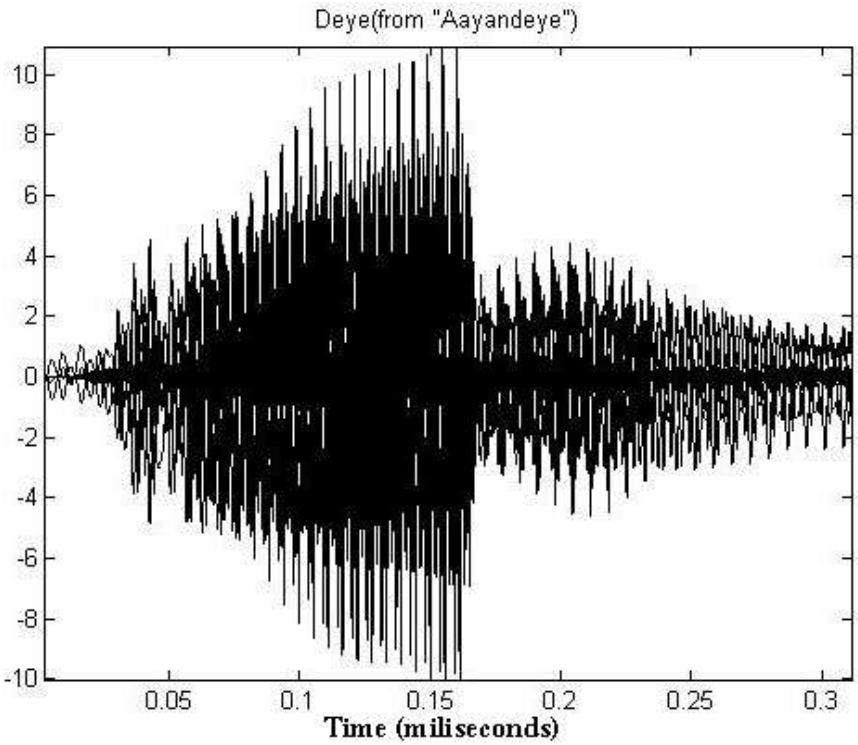
- صامت‌ها به ترتیب حروف الفبا باشند و مصوت‌ها نیز بـه ترتیـب " َ ، ِ ، ُ ، آ، ای، او" باشـند. مثلا در ترکیب CV حروف الفبا به ترتیب ابتدا با کسره چیده مـی‌شـوند و سـپس بـا فتحـه و ضمه و .... در ادامه، ترکیب‌های VC و CC و ... می‌آیند.
- صامت‌ها و مصوت‌ها با توجه به تعداد تکرارشان در متون مختلف طبقه‌بندی مـی‌شـوند. مـثلا حروف صدادار و "م"، "ن" و ... بیشترین کاربرد را دارند و در این لیست در اولویت قرار دارند. با این وجود به اطلاعات آماری از تعداد صامت‌ها و مصوت‌ها در متون مختلف زبان فارسی نیاز داریم.
- دایفون‌ها را از بین ۲۰۰۰ یا ۳۰۰۰ واژه‌ی پرکاربرد فارسی اسـتخراج مـی‌کنـیم و بـه ترتیـب بیشترین تکرار، آنها را کنار هم قرار می‌دهیم.

برای چیدمان دایفون‌های صوتی و متنی از یک نوع کد استفاده می‌شود. مثلا اگر برای دایفون "تِ" از کد ۰۲۳ استفاده کنیم آنگاه برای هرگونه از حالات گفتار یک رقم در سمت راست قرار می‌دهیم. در این صورت کدهای دایفون "تِ" شامل ۰۲۳۱، ۰۲۳۲، ۰۲۳۳، ۰۲۳۴، ۰۲۳۵، ۰۲۳۶، ۰۲۳۷ می‌باشد.

۴-۱۰-۲. هنجارسازی سنتز دایفون‌ها

بعضی مشکلات در هنگام سنتز گفتار وجود دارند که گوش دادن به خروجی آن(هرچند طبیعی باشد) باعث ناخوشایند بودن صدا می‌شوند. برخی از این مشکلات و راه‌حل‌های آن‌ها را می‌توان به موارد زیر تقسیم کرد:

- شدت: معمولا فرد گوینده، یکنواختی شدت گفتار را نمی‌تواند به طور کامل رعایت کند. این اتفاق باعث می‌شود قطعات پس از تقطیع از نظر شدت، با یکدیگر اختلاف داشته باشند. با سنتز این قطعات، در کنار هم، اختلاف شدت آن‌ها یقینا شنیده شده و خودنمایی می‌کند. در این صورت پیشنهاد می‌شود که پس از تقطیع، تک تک دایفون‌ها، از نظر شدت گفتار تصحیح شوند. در شکل ۴-۱۲ دو قطعه دایفون بدون هنجارسازی، سنتز شده‌اند.



شکل ۴-۱۲: اختلاف شدت بین دو قطعه‌ی سنتز شده، یکی از مشکلات سنتز و قابل حل در مرحله‌ی هنجارسازی

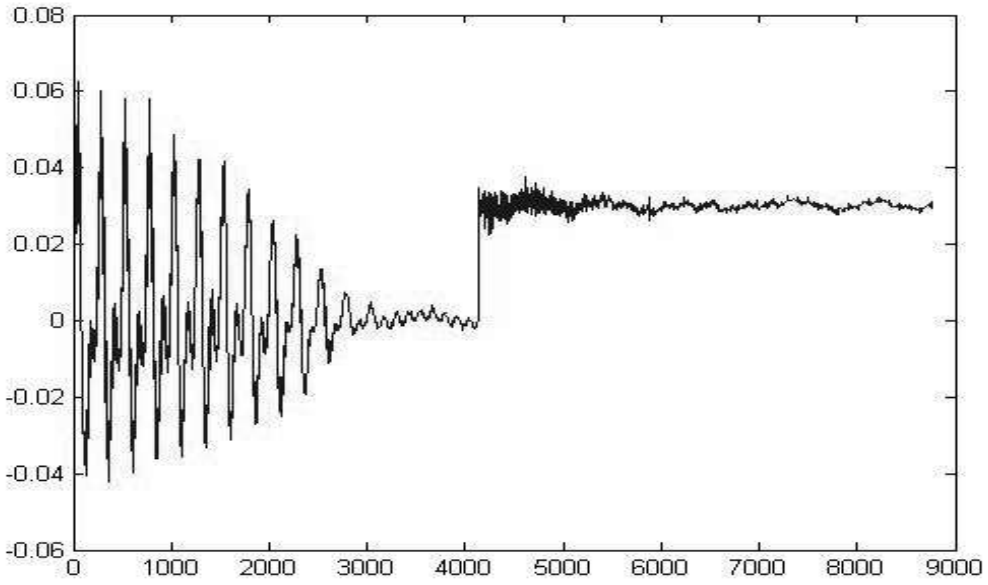
اگر بعد از استخراج دایفون‌ها، این تصحیح رخ نداد، می‌توان در زمان سنتز، طبق مدل‌های ریاضی و آماری، این مشکل را برطرف نمود. باید توجه داشت، چون طول زمانی دایفون‌ها، متفاوت هستند، باید ابتدا آن‌ها را نرمالیزه کرد و سپس به کار برد.

- اختلاف فاز: این مشکل زمانی اتفاق می‌افتد که دو قطعه در کنار هم قرار می‌گیرند و نمونه‌ی آخر قطعه‌ی اول و نمونه‌ی اول قطعه‌ی بعدی، اختلاف فاز زیادی داشته باشند. برای رفع این مشکل می‌توان در هنگام استخراج، نمونه‌های ابتدایی و انتهایی را از نظر شدت انرژی با روش‌های مختلف، کم کرد. سپس نقطه‌ی شروع سیگنال را از صفر و هنگامی که شدت انرژی مثبت می‌شود انتخاب می‌کنیم و پایان سیگنال نیز صفر است و در جایی است که شدت از منفی به صفر می‌رسد. در شکل ۴-۱۳ مشخص شده است که می‌توان انتهای دایفون "زَ" را روی خط قرمز گرفت. یعنی جایی که شدت سیگنال از مقادیر منفی به صفر می‌رسد را به عنوان انتهای سیگنال در نظر می‌گیریم. ابتدای سیگنال که حرف "ز" است، فرکانس بالایی دارد ولی به صورت تقریبی می‌توان نقطه‌ی شروع سیگنال از صفر را مشخص کرد. این ویرایش می‌تواند در نرم‌افزارهای مختلف ویرایش صدا، مانند نرم‌افزار WavePad Sound Editor Master Edition که نگارنده از آن بهره برده است، انجام شود.

شکل ۴-۱۳: دایفون "زَ" و حل مشکل اختلاف فاز با استفاده از نرم‌افزار ویرایش صدای WavePad Sound Editor (Master Edition)

همچنین می‌توان این مشکل را با مدل‌های ریاضی برطرف کرد. به این صورت کـه سـیگنال را در توابع نمایی میراشونده ضرب کنـیم تـا ابتـدا و انتهـای آن دارای شـدت کمـی باشـند و اختلاف فاز محسوس نباشد. البته در این روش سیگنال اصلی دچار اعوجاج مـی‌شـود و باعـث خرابی سیگنال، و در نهایت خرابی سیگنال سنتز شده نهایی می‌شود.

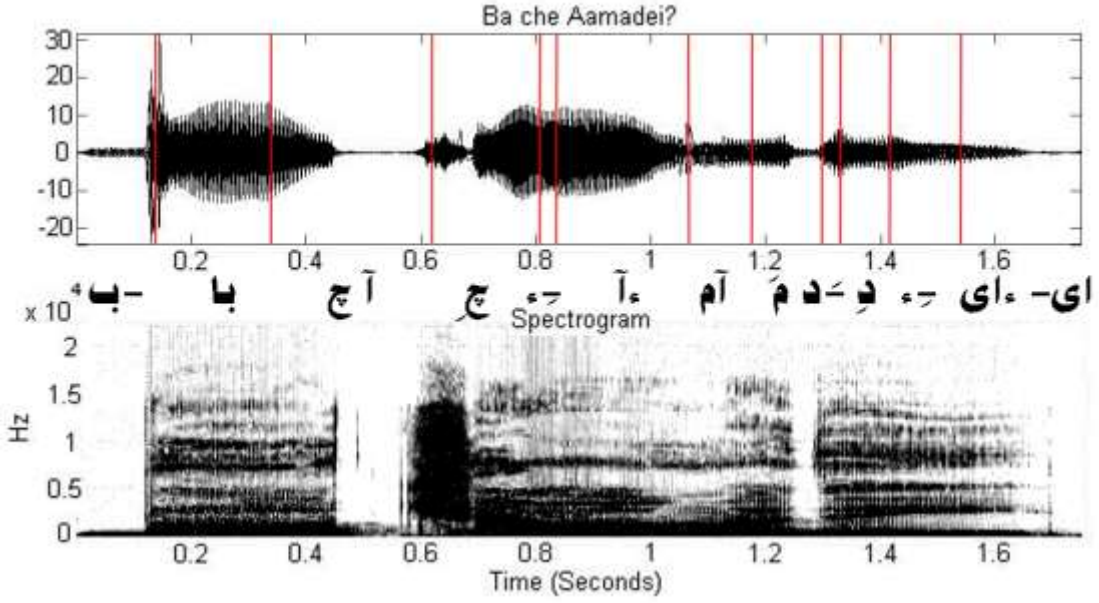
- میانگین: در زمان استخراج باید در نظر داشت که تمام سیگنال‌ها میانگین زمانی صفر داشـته باشند. این اشکال زمانی رخ می‌دهد که هنگام ضبط گفتار، امکان دارد مقدار ثـابتی بـه تمـام نمونه‌های ضبط شده اضافه گردد که با صفر کردن میانگین، ایـن مشـکل برطـرف مـی‌گـردد. شکل ۴-۱۴ به هم پیوستن دو قطعه را نشان می‌دهد که دارای میـانگین‌هـای متفـاوت مـی-باشند.



شکل ۴-۱۴: اختلاف میانگین بین دو قطعه‌ی سنتز شده، یکی از مشکلات سنتز و قابل حل در مرحله‌ی هنجارسازی

مکث: همان‌گونه که در فصل سوم قسمت زبان‌شناسی به بحث مکث یـا درنـگ پرداختـه‌ایـم، این موضوع در قسمت گفتار نیز مورد استفاده است. به طور کلی می‌توان دو نـوع مکـث را در نظر گرفت: "مکث میان‌کلمه‌ای" و "مکث میان‌هجایی". مکـث میـان‌کلمـه‌ای از نظـر زمـانی

بیشتر از مکث میان‌هجایی است. این نکته باید در زمان سنتز مد نظر قرار گرفته شود، وگرنه سنتز خروجی نامفهوم خواهد شد. برای حل این مشکل، نگارنده حالت هفتم دایفونی خود را پیشنهاد داده است که همان دایفون‌های میان‌واژه‌ای هستند. برای حل مشکل مکث میان-هجایی نیز می‌توان از سکوت کوتاهی بین دایفون‌ها استفاده کرد که چون این فواصل هیچ‌گاه ثابت نیستند، نگارنده ترجیح داده است که از این فاصله، صرف نظر کند.



شکل ۴-۱۵: فاصله‌های کوچک، فاصله‌ی میان‌واژه‌ای و فاصله‌های بزرگ، فاصله‌ی میان‌کلمه‌ای

با توجه به آن‌چه در شکل ۴-۱۵ دیده می‌شود، اثبات شده است که اولا فاصله‌های میان-هجایی و میان‌واژه‌ای، مقدار ثابتی نیست و ثانیا نمی‌توان تعریف درستی از این فاصله داشت و روش خاصی برای استخراج این فواصل وجود ندارد.

۴-۱۰-۳. محاسبه‌ی حافظه‌ی اشغالی در تولید پایگاه دایفونی با استفاده از روش پیشنهادی

در این بخش می‌خواهیم محاسبه‌ی تقریبی روی حجم دادگان متنی و گفتاری انجام دهیم. اطلاع از مقدار حجم آن‌ها از این رو مفید است که در انتخاب سخت‌افزار و سرعت پردازنده‌ی سیستم راه‌گشا باشد.

داده‌های متنی

در فصل دوم درباه‌ی دادگان مورد نیاز برای یک سیستم متن به گفتار توضیح داده شده است. در این قسمت می‌خواهیم حجم دادگان مورد نیاز را تخمین بزنیم. و از مطالب ذکر شده در فصل سوم می‌دانیم، از ترکیب سکون و ۲۳ صامت و ۶ مصوت (با ساختار C-، -C، CV، VC، CC و -V)، حدود ۸۵۰ نوع دایفون خواهیم داشت.

در یک سیستم متن به گفتار دو نوع پایگاه داده برای پردازش متن استفاده می‌شود. یکی پایگاه داده‌ای شبیه به پایگاه داده بی‌جن‌خان است که در آن نقش‌های دستوری ممکن برای کلمات در جمله ذخیره شده است و دیگری پایگاه داده‌ای برای ذخیره‌سازی دایفون‌ها و هجاهای همان واژگان. پایگاه داده‌ی متنی دوم شامل تمام واژگان پیکره بی‌جن خان است که در آن به جای مشخص کردن نقش دستوری کلمات، تقسیم بندی دایفونی و هجایی آن‌ها را ذخیره می‌کنیم. این تقسیم بندی بهتر است با اعداد انجام شود که برای برنامه نویسی و فراخوانی آن‌ها دستمان بازتر باشد[1].

اگر تعداد واژگان رایج در زبان فارسی را ۴۰۰۰۰ واژه در نظر بگیریم و هر کلمه را دارای ۵ واج، که تعداد دایفون‌های آن به طور میانگین ۱۰ عدد، و تعداد هجاهای هر واژه را ۳ عدد در نظر بگیریم، به ۲ مرز هجایی نیاز خواهیم داشت. با در نظر گرفتن حافظه‌های دو بایتی برای هر یک از واج‌ها، دایفون‌ها و مرز هجاها، به محاسبات زیر می‌رسیم:

۵ واج+ ۱۰ دایفون+ ۲ مرز هجایی= ۱۷

۲×۱۷ بایت= ۳۴ بایت

۳۴× ۴۰۰۰۰ واژه= ۱۳۶۰۰۰۰ بایت= ۱/۲۹ مگابایت

[1] شماره‌گذاری دایفون‌ها باید با اعداد سه رقمی انجام شود و از آنجا که تعداد دایفون‌ها ۸۵۰ عدد تخمین زده شده است برای نمایش مرز هجایی می‌توان از عدد ۸۸۸، ۹۹۹، یا ۰۰۰ استفاده کرد. مثلا کلمه کلاس به صورت –ک کـِ لـِل لا آس و کدگذاری آن به صورت فرضی ۱۵۰۲۳۰۰۰۰۳۰۲۱۰۲۲۰۵ می‌باشد.

پس برای ذخیره‌سازی اطلاعات دایفونی هر واژه به طور میانگین به ۳۴ بایت نیاز داریم و در نتیجه برای ۴۰۰۰۰ واژگان فارسی، ۱/۲۹ مگابایت لازم است، که این مقدار در مقایسه با سیستم‌های موجود امروزی مقدار قابل توجهی نیست.

۴-۱۰-۳-۲. داده‌های صوتی

یک کیفیت عالی برای ضبط صدا فرکانس نمونه‌برداری ۱۶ کیلوهرتز و عمق حافظه ۱۶ بیت است. برای کاهش حجم حافظه جهت ذخیره‌سازی می‌توان از کاهش کیفیت سیگنال، از طریق کاهش فرکانس نمونه‌برداری از ۱۶ کیلوهرتز به ۸ کیلوهرتز و یا کم کردن عمق حافظه برای ذخیره‌ی نمونه‌ها از ۱۶ بیت به ۸ بیت، حجم حافظه‌ی اشغالی را در هر کدام از این دو حالت به نصف، و یا در صورت رعایت هر دو مورد، حجم مورد نیاز را به یک چهارم مقدار اولیه رساند. البته آزمایشات و تجربیات نگارنده نشان داده است که فرکانس نمونه‌برداری ۱۶ کیلوهرتز و عمق حافظه‌ی ۸ بیت (۱۲۸ کیلوبیت بر ثانیه)، برای این منظور مناسب‌تر است.

اگر فرض شود که ۸۵۰ دایفون در نهایت نیاز داریم و متوسط طول هر دایفون شامل ۵۰۰۰ نمونه باشد با توجه به محاسبات زیر می‌توان حجم نهایی فایل‌های صوتی را تقریب زد:

هفت حالت پیشنهادی برای دایفون× ۸۵۰ تعداد کل دایفون‌ها= ۵۹۵۰ تعداد کل دایفون‌ها برای هفت حالت

۵۹۵۰× ۵۰۰۰ نمونه× ۲ بایت برای هر نمونه = ۵۹۵۰۰۰۰۰ بایت= ۵۶/۷۴ مگابایت

همانطور که ملاحظه می‌شود کل حجم مورد نیاز برای دادگان صوتی تقریبا ۵۷ مگابایت می‌باشد، که نسبت به حجم حافظه‌ی سیستم‌های موجود، بسیار کم‌تر است.

فصل پنجم:

نتایج و تحلیل شبیه‌سازی‌ها

۵-۱. مقدمه

در این فصل می‌خواهیم روش پیشنهادی را با سایر روش‌های مشابه بسنجیم و کارکرد آن را در قالب چند مثال نشان دهیم. چون در روش پیشنهادی، سعی شده است که آهنگ جملات و تکیه‌ی کلمات به خوبی مدل‌سازی شوند، باید از نظر صحت نوای گفتار امتیاز بسیار خوبی داشته باشد. ولی به دلیل عدم وجود سیستم هنجارسازی سنتز در این پروژه، بی‌شک در آزمایشات انجام شده از نظر روانی بیان انتظار نمره‌دهی چندان خوبی توسط شنوندگان نداریم. البته یک سیستم قوی سنتز گفتار شامل بخشی با قابلیت هنجارسازی سنتز می‌باشد که باعث روانی گفتار خروجی می‌شود.

با این مقدمه آزمایشات و مقایسات را انجام می‌دهیم و در فصل نتیجه گیری بحث بیشتری در مورد نتیجه آزمایشات صورت می‌گیرد.

۵-۲. سنتز گفتار با استفاده از روش پیشنهادی

با توجه به آن‌چه در فصل روش پیشنهادی گفته شده است، برای تهیه‌ی دادگان دایفونی و استفاده آن در سیستم سنتز گفتار، هفت مدل دایفون برای ایجاد حالت‌های مختلف گفتاری پیشنهاد شده است.

به همین منظور، برای هر کدام از انواع جمله از نظر حالت گفتاری (که در فصل روش پیشنهادی درباره‌ی آن توضیح داده شده است) یک جمله‌ی کوچک و ساده بازسازی شده است. چهار جمله با حالت‌های سوالی با واژه "آیا"، خبری و سوالی ساده، تعجبی و سوالی-تعجبی در قسمت سنتز بازسازی شده‌اند:

- "آینده‌ی خوبی داری؟"
- "چگونه پیوست؟"
- "چه آوازی می‌خواند!"
- "به چه سرعتی می‌سازد؟!"

همچنین این جملات، در قسمت ضبط توسط گوینده بیان شده است و در ادامه، سیگنال صوت آن‌ها را از دیدگاه‌های مختلف با هم مقایسه می‌کنیم.

۵-۲-۱. دایفون‌های استفاده شده در جمله‌ی پرسشی با واژه‌ی "آیا"

جمله‌ی "آینده‌ی خوبی داری؟" را در خروجی سیستم سنتز تولید می‌کنیم. تکیه‌ی کلمات "آینده‌ی خوبی" در هجاهای "ده" و "بی" وجود دارد. دایفون‌هایی که باید برای این هجاها استفاده کرد به صورت زیر است:

- هجاهای "آینده‌ی خوبی" شامل دایفون‌های تکیه‌دار زیر خواهد بود:

  ده = ن د (دایفون CC) + دِ (دایفون CV)

  بی = اوب (دایفون VC) + بی (دایفون CV)

- هجاهای جمله‌ی مذکور شامل دایفون‌های بدون تکیه‌ی زیر خواهد بود:

  آ = -ء (دایفون C-) + ءآ (دایفون CV)

  یَن = آی (دایفون VC) + یَ (دایفون CV) + نَ (دایفون VC)

  یِ = یِ (دایفون VC) + یِ (دایفون CV)

  خو = خو (دایفون CV)

- دایفون‌های میان‌واژه‌ای در جمله‌ی مذکور به صورت زیر می‌باشد:

  خِ (دایفون میان واژه‌های "آینده‌ی خوبی")

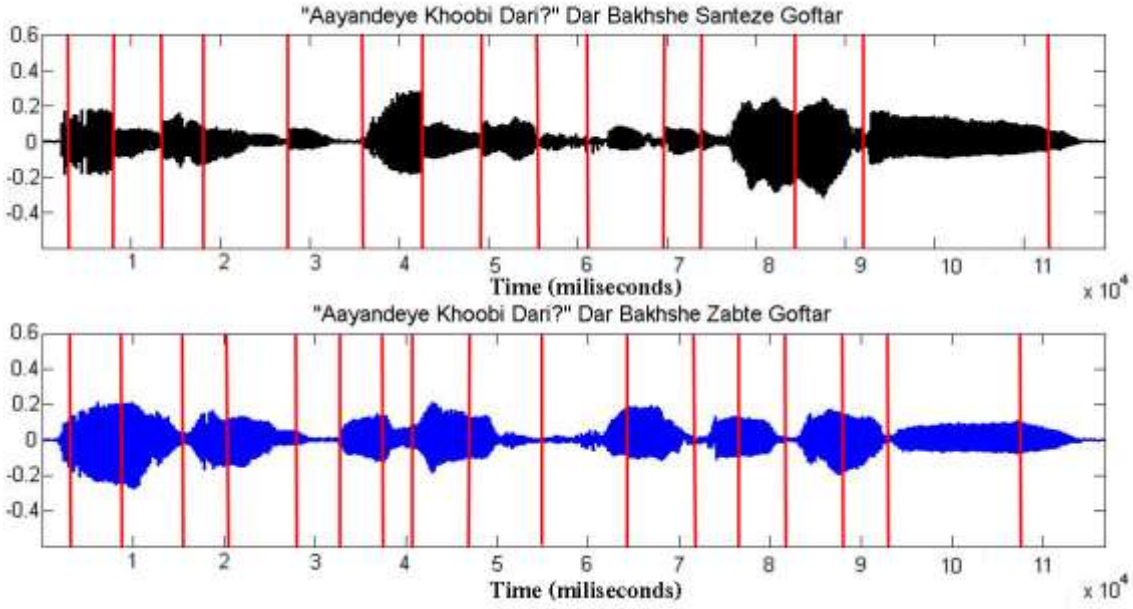
  ای د (دایفون میان واژه‌های "خوبی داری؟")

- دایفون‌های فعل جمله‌ی مذکور:

  دا (دایفون استخراجی از فعل سوالی که هجای ابتدایی آن، "دا" باشد)

  ری = آر + ری + ای- (دایفون استخراجی از فعل سوالی که هجای انتهایی آن، "ری" باشد)

در شکل ۵-۱ سیگنال‌های موج این جمله را نشان می‌دهد، شکل بالایی سیگنال خروجی سیستم سنتز است و شکل پایینی سیگنال ضبط شده توسط نگارنده در هنگام ضبط است، که جمله را بدون وقفه خوانده است.



شکل ۵-۱: مقایسه‌ی سیگنال‌های جمله‌ی "آینده‌ی خوبی داری؟" خروجی سیستم سنتزشده (سیگنال بالا) و ضبط شده توسط گوینده (سیگنال پایین)

همانطور که در شکل بالا روشن است، سیگنال ضبط شده توسط نگارنده یکنواخت و بدون شدت‌های ناگهانی است، در صورتی که در سیگنال سنتز شده این‌گونه نیست.

نکته‌ی مهم دیگر این است که دایفون‌های سیگنال گفتار طبیعی و مصنوعی بالا دارای طول زمانی متفاوتی هستند. پس باید هنگام تهیه‌ی دادگان به این قضیه دقت کرد که دایفون مورد نیاز را از کدام هجا و کلمه‌ای استخراج کنیم که در سیستم سنتز دچار مشکل نشویم.

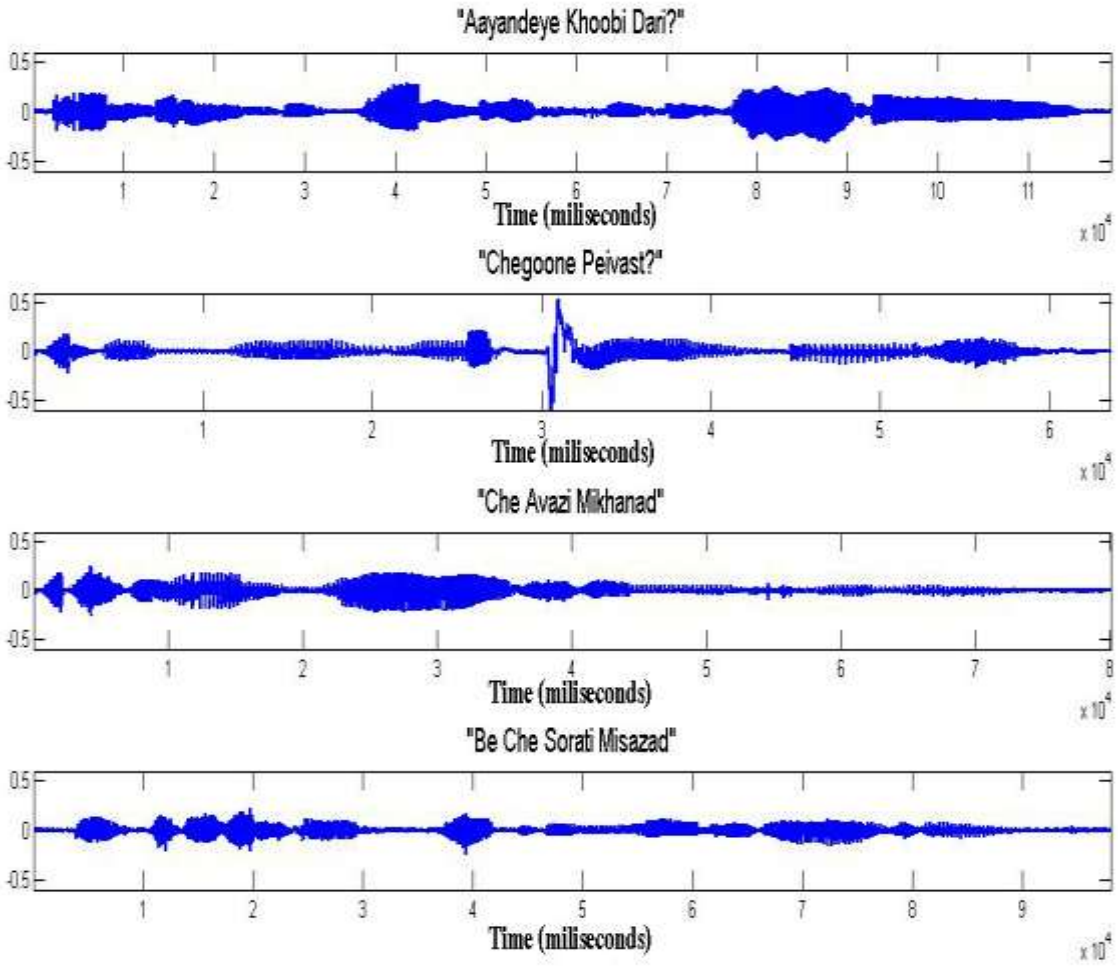
۵-۲-۲. تولید جملات با حالت‌های مختلف گفتاری

برای شبیه‌سازی روش پیشنهادی بیش از ۴۰ جمله‌ی کوتاه توسط نگارنده ضبط شد و دایفون‌های مورد نیاز برای تولید جمله‌های مورد نظر، استخراج شدند. با استفاده از نرم افزار WavePad Sound

Editor (Masters Edition) عملیات استخراج دایفون‌ها و سنتز آن‌ها قرار گرفت. هم‌چنین در محیط "متلب" نیز عمل سنتز انجام شد، که با مقایسه‌ی آن با سنتز "نرم‌افزار ویرایش صدا" به این نتیجه رسیدیم که سنتز در محیط "نرم‌افزار ویرایش صدا"، خروجی بهتری خواهد داشت. به این علت که در هنگام اتصال دو تکه‌ی گفتار به هم، نرم‌افزار قابلیت نرم‌سازی سیگنال‌های پیوسته شده را دارد و این اتفاق باعث می‌شود کوبه‌ای بودن سنتز خروجی کمتر به نظر بیاید.

چهار جمله با حالت‌های مختلف گفتاری در قسمت سنتز تولید کردیم و می‌خواهیم کیفیت سنتز آن‌ها را با یک روش مناسب و پیشرفته، در قسمت بعد، مقایسه کنیم.

در شکل ۵-۲ شکل موج سیگنال این جملات رسم شده است.



شکل ۵-۲: شکل موج جملات سنتز شده با حالت‌های مختلف گفتاری

در شکل ۵-۲ همانطور که دیده می‌شود، سنتزهای انجام‌شده دارای مشکلاتی است که به بررسی برخی از آن‌ها در فصل روش پیشنهادی پرداخته‌ایم. مثلا مشکل اختلاف شدت گفتار بین دایفون‌های مختلف، درنگ‌های میان‌هجایی و ... است، که برای بیشتر آن‌ها راه حل‌هایی پیشنهاد کرده‌ایم.

۵-۳. مقایسه‌ی روش پیشنهادی با روش مبتنی بر دایفون‌های تکیه دار و بدون تکیه

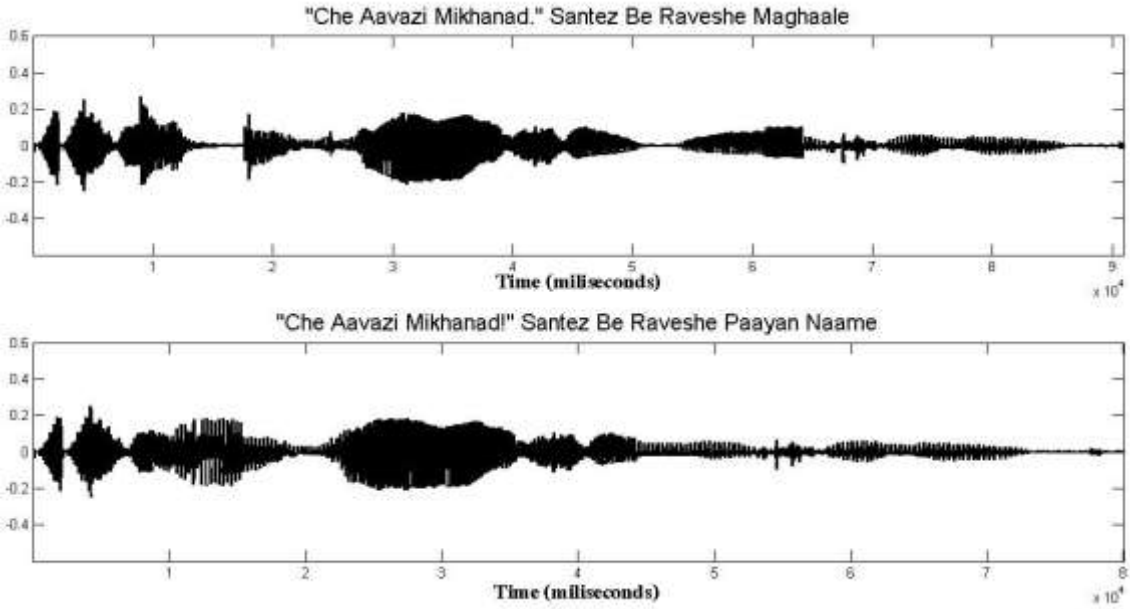
مقاله‌ی [۱۹] روشی را پیشنهاد کرده است که شبیه به روش پیشنهادی پایان‌نامه می‌باشد. در آن‌جا پیشنهاد تهیه‌ی دو نوع دادگان، دادگان هجایی و دادگان دایفونی داده شده است. به دلایل بیان شده در مقدمه‌ی فصل روش پیشنهادی، این مقاله دادگان دایفونی را بسیار مناسب‌تر از دادگان هجایی می‌داند. با این وجود که این مقاله طرز تهیه و به انجام رساندن دو دادگان هجایی و دایفونی را به طور کامل توضیح داده است، ولی برای یک سیستم سنتز، دادگان دایفونی را ترجیح داده است.

این مقاله، طرز تهیه‌ی یک دادگان دایفونی را از جهات مختلف توضیح داده است. در آن‌جا فقط دایفون‌ها را از لحاظ تکیه‌دار بودن یا نبودنشان استخراج کرده است. با توجه به توضیحات مقاله، نگارنده بر آن شد که طبق نظریه‌ی آن، یک جمله تولید کند و با نتیجه‌ی روش پیشنهادی در این پایان‌نامه مقایسه کند. برای این کار فقط لازم بود در جمله‌ی مذکور، هجاهای تکیه‌دار و بدون تکیه را مشخص کنیم، سپس در بین جملات ضبط شده، به دنبال دایفون‌های مورد نظرمان بگردیم و آن‌ها را استخراج کنیم. در شکل ۵-۳ این مقایسه انجام شده است. سیگنال بالایی، حاصل از سنتز به روش مقاله است و سیگنال پایینی، حاصل از سنتز به روش پیشنهاد شده در پایان‌نامه است.

در تحلیل اختلاف این دو سیگنال، می‌توان به نکات زیر اشاره کرد:

- در روش پیشنهادی به دلیل وجود دایفون‌های میان‌واژه‌ای، پیوستگی گفتار و طبیعی بودن آن بهتر رعایت شده است.

- در روش پیشنهادی فعل جمله را از نمونه‌های فعـل تعجبـی اسـتخراج کـردیم ولـی در روش مقاله در هجای اول از دایفون‌های تکیه‌دار استفاده شده است. در این مقایسه کـاملا مشـخص است که حالت فعل تعجبی در روش پایان‌نامه به درستی بیان شده است ولـی در روش مقالـه جمله به صورت خبری بیان می‌شود.
- با توجه به نکات گفته شده می‌توان نتیجه گرفت که جمله‌ی سنتز شـده بـا روش پیشـنهادی حالت بسیار طبیعی‌تری نسبت به روش [۱۹] دارد.



شکل ۵-۳: شکل موج بالایی، سیگنال سنتز شده به روش مقاله [۱۹] و پایینی به روش پیشنهادی پایان‌نامه

۵-۴. مقایسه‌ی روش پیشنهادی با روش‌های موجود

با اشراف بر این موضوع که مقالات ارائه شده در این موضوع و درباره‌ی زبان فارسی تعداد کمـی دارنـد، عملکرد روش پیشنهادی پایان‌نامه را نیز با چند مقالـه‌ای کـه نتـایج را بـا معیـار MOS سـنجیده‌انـد، مقایسه می‌کنیم. در این روش، جملات سنتز شده را برای ۱۰ مرد و ۱۰ زن، دارای میانگین سـنی ۲۰ تا ۳۵ سال، فارسی‌زبان و بدون سابقه‌ی اختلال شنوایی، پخش می‌کنند و از نظر "صحت نوای گفتار"،

"قابل فهم بودن" و "روانی بیان" از ۱ تا ۵ نمره می‌دهند (بهترین نمره ۵ است). این روش ارزیابی، روش دقیقی نیست و نمی‌توان شرایط آن‌ها را یکسان کرد و در کل شاید دارای خطای ارزیابی باشد، ولی چون هنوز روش مصوب و دقیقی برای مقایسه‌ی گفتار سنتز شده ارائه نشده است، به نتایج این ارزیابی بسنده می‌کنیم.

در [۱۹] که سال ۱۳۸۸ ارائه شده است، در مورد مراحل و نحوه‌ی تهیه‌ی دادگان‌های هجایی و دایفونی زبان فارسی برای استفاده در سیستم تبدیل متن به گفتار توضیح داده است. ولی در آن در مورد ارزیابی و مقایسه با روش‌های دیگر سخنی به میان نیاورده است. همچنین در مقاله‌ی [۲۱] که سال ۱۳۸۹ در مورد طراحی و پیاده‌سازی دادگان دایفونی زبان فارسی تحقیق کرده است، درباره‌ی مقایسه و ارزیابی بحثی به میان نیاورده است.

در مقاله‌ی [۲۲] سال ۱۳۸۴ در مورد طراحی و پیاده‌سازی سیستم تبدیل متن به گفتار طبیعی FTTS[1] توضیح داده است که میانگین نمره‌ی آن در آزمون MOS، ۳/۵۹ می‌باشد.

در مقاله‌ی [۱۴] که در سال ۱۳۸۹ به چاپ رسیده است. در این مقاله ابتدا هجاهای غیرنوایی کلمات هم‌گذاری شده به روش PSOLA مورد پردازش قرار گرفتند تا به گفتار طبیعی نزدیک شوند و سپس با استفاده از واحدهای حساس به بافت نوایی هم‌گذاری انجام شده و به گفتار طبیعی نزدیک - می‌شود. در آن‌جا با مقایسه دو روش بیان شده به این نتیجه رسیده است که واحدهای حساس به بافت نوایی در سیستم سنتز، عملکرد بهتری دارند و میانگین آزمون MOS آن‌ها بین ۳/۵ تا ۴/۵ است.

در مقاله‌ی [۱] سال ۱۳۹۳ به زبان انگلیسی منتشر شده است. درباره‌ی روش پیشنهادی این مقاله در فصل دوم توضیح داده شده است. در این مقاله، عملیات استخراج واحدهای گفتار با استفاده از سکوت‌های بین دو یا سه هجا است که نام این واحدهای گفتار را SBS گذاشته‌اند. در این مقاله برای مقایسه و ارزیابی، از آزمون MOS استفاده کرده است و روش خود را با دو روش دیگر، مقایسه کرده است. یکی روش درخت تصمیم‌گیری و دیگری روش استفاده از ترایفون می‌باشد.

[1] *Farsi Text-to-Speech*

در جدول ۵-۱ امتیاز شنوندگان به سنتز حاصله با روش پیشنهادی را نشان می‌دهد. این جدول با پخش چهار جمله‌ی سنتز شده و امتیاز کلی شنوندگان به آن‌ها از نظر قابل فهم بودن، صحت نوای گفتار و روانی بیان، تهیه شده است.

جدول ۵-۱: امتیاز شنوندگان به روش پیشنهادی پایان‌نامه با استفاده از معیار MOS

شنوندگان	قابلیت فهم	صحت نوای گفتار	روانی بیان
مرد۱	۴/۷۵	۴/۲۵	۴/۲۵
مرد۲	۴/۷۵	۴/۲۵	۳
زن۱	۳/۷۵	۳/۷۵	۳/۷۵
مرد۳	۴/۵	۳/۷۵	۳
مرد۴	۵	۵	۳/۷۵
زن۲	۵	۵	۴/۷۵
مرد۵	۵	۴/۵	۲/۷۵
مرد۶	۴/۵	۳/۷۵	۳/۵
زن۳	۴/۷۵	۳/۷۵	۲/۷۵
مرد۷	۴/۷۵	۴/۵	۴
زن۴	۴/۷۵	۴/۲۵	۴/۵
مرد۸	۴/۲۵	۴/۲۵	۳/۵
زن۵	۴/۲۵	۴	۳
زن۶	۵	۵	۴
زن۷	۵	۵	۳/۵
زن۸	۵	۴/۵	۴/۵

مرد۹	۵	۵	۵
مرد۱۰	۵	۴/۷۵	۴
زن۹	۵	۴/۷۵	۳/۵
زن۱۰	۵	۴/۲۵	۳/۵
میانگین	۴/۷۵	۴/۴۱	۳/۷۳

در جدول ۵-۲ امتیاز شنوندگان به سنتز با روش مقاله [۱۹] را نشان می‌دهد. این جدول با پخش کردن سنتز حاصل از جمله‌ی "چه آوازی می‌خواند" و امتیاز شنوندگان حاصل شده است. همانطور که در فصل روش پیشنهادی توضیح داده شده است، این روش فقط تکیه‌دار بودن یا نبودن هجا را در نظر می‌گیرد و به حالت‌های دیگری که در این پایان‌نامه به آن‌ها اشاره شده است، بی‌توجه بوده است.

جدول ۵-۲: امتیاز شنوندگان به روش شبیه‌سازی شده دایفون‌های تکیه‌دار وبدون تکیه با استفاده از معیار MOS

شنوندگان	قابلیت فهم	صحت نوای گفتار	روانی بیان
مرد۱	۳	۲/۵	۲/۵
مرد۲	۵	۳	۲/۵
زن۱	۲	۲	۲
مرد۳	۴	۲/۵	۲
مرد۴	۵	۳	۳/۵
زن۲	۵	۳	۳
مرد۵	۵	۴/۵	۲
مرد۶	۳	۲	۲

زن۳	۳	۲	۲
مرد۷	۳	۳	۲
زن۴	۵	۳	۳/۵
مرد۸	۳	۴	۲
زن۵	۵	۴	۲/۵
زن۶	۵	۴	۴
زن۷	۵	۴	۳
زن۸	۵	۳	۳/۵
مرد۹	۵	۴	۳/۵
مرد۱۰	۵	۳	۳
زن۹	۴	۳	۲
زن۱۰	۵	۳	۲
میانگین	۴/۲۵	۳/۱۳	۲/۶۳

در جدول ۳-۵ مهم‌ترین روش‌ها و هم‌چنین روش پیشنهادی نگارنده‌ی پایان‌نامه را با استفاده از آزمون MOS مقایسه کرده‌ایم. نتایج این آزمون‌ها در [۱] آمده است. در این مقاله روش SBS برای واحدهای گفتار در نظر گرفته شده است (که در فصل دوم درباره‌ی آن توضیح داده شده است) و روش خود را با روش‌های موجود مورد مقایسه قرار داده است. در این‌جا ما نیز روش خود را با جدول تهیه شده در این مقاله مقایسه می‌کنیم.

جدول ۵-۳: مقایسه‌ی روش‌های دیگر موجود در مقاله [۱] با روش پیشنهادی و روش شبیه‌سازی شده در معیار MOS

روش استفاده شده	قابلیت فهم	صحت نوای گفتار	روانی بیان
درخت تصمیم‌گیری	۴/۲	۴/۴	۴/۱
استفاده از ترایفون	۳/۸	۳/۹	۳/۵
روش SBS	۴/۸	۳/۶	۳/۸
دایفون‌های تکیه دار و بدون تکیه	۴/۲۵	۳/۱۳	۲/۶۳
روش پیشنهادی پایان‌نامه	۴/۷۵	۴/۴۱	۳/۷۳

با دقت در امتیازات متوجه می‌شویم که قابلیت فهم و صحت نوای گفتار روش پیشنهادی بسیار بالا و مطلوب است. ولی روانی بیان آن نمره‌ی معمولی دارد که آن نیز می‌تواند بهتر شود، به این صورت که در فصل روش پیشنهادی درباره‌ی هنجارسازی دایفون‌ها توضیح مفصلی دادیم. یکی از کارهای مهمی که در تهیه‌ی پایگاه داده بسیار مهم می‌باشد این است که شدت واحدهای گفتار پس از استخراج در یک سطح خاص قرار گرفته بشوند و این کار بسیار تخصصی است و باید در استودیو و به وسیله‌ی افراد باتجربه انجام شود. پس نتیجه می‌شود که روش پیشنهادی می‌تواند از نظر روانی بیان نیز دارای امتیاز بسیار خوبی باشد.

با تهیه‌ی دادگان دایفونی به روش پایان‌نامه، حجم بسیار اندکی نسبت به دادگان هجایی اشغال خواهد شد. پس می‌توان گفت سرعت جستجو و فراخوانی واحد نیز، در این روش بسیار بالاتر از روش-هایی است که از هجا به عنوان واحد گفتار استفاده می‌کند. تا کنون دادگانی که اجزای آن دایفون باشند به طور کامل ارائه نشده است. با مراجعه به سایت دادگان فارسی[1] می‌توان در مورد این موضوع، تحقیق بیشتری کرد.

[1] http://www.dadegan.ir/

فصل ششم:

نتیجه‌گیری و پیشنهادات

۶-۱. نتیجه‌گیری

برای انجام این پایان‌نامه، می‌بایست درباره‌ی علوم پایه‌ی زیر، آشنایی خوبی کسب می‌کردیم:

- زبان‌شناسی و علم تکیه، آهنگ و مکث در جمله
- علم رایانه و کار با نرم افزار MATLAB و نرم‌افزارهای ویرایش صدا

با داشتن این زمینه‌های علمی، پروژه انجام شد. در ادامه دایفون‌ها را برای استفاده در ورودی یک سیستم سنتز گفتار مناسب دیدیم و با تحقیقاتی که انجام دادیم به این نتیجه رسیدیم که هفت مدل دایفون می‌توان استخراج کرد که با این کار بتوانیم گفتار بهتری را در خروجی سیستم داشته باشیم.

روش پیشنهادی با بهترین روش‌های موجود مقایسه شد. یکی از روش‌های موجود را در شرایط یکسان با روش پیشنهادی، شبیه‌سازی کردیم و توسط شنوندگان امتیازدهی شد. روش پیشنهادی از لحاظ درک مطلب و صحت نوای گفتار، بهترین نمره را گرفت ولی در روانی بیان نتوانست امتیاز خیلی بالایی دریافت کند. برای توجیه بالا نبودن امتیاز روانی بیان خروجی نیز، دلایلی بیان کردیم که از جمله‌ی آن می‌توان به الزام یکسان کردن انرژی دایفون‌ها در پایگاه داده اشاره کرد.

در نهایت نتیجه می‌شود که پیشنهاد انواع دایفون و تعدد آن در یک پایگاه داده‌ی دایفونی باعث می‌شود که مدل‌های بیشتری از آهنگ گفتار و تکیه‌ی کلمات، در سیستم سنتز گفتار، پیاده‌سازی شود. در نتیجه به جای این‌که نوای گفتار به وسیله‌ی مدل‌های پیچیده‌ی آماری و ریاضی، بر روی جمله‌ی سنتز شده اعمال شود، می‌توان با استفاده از انواع دایفون‌های موجود در دادگان و جاسازی دایفون مناسب در جایگاه خودش، نوای گفتار طبیعی را، بدون استفاده از مدل‌های آماری و ریاضی تولید کرد. این روش نه تنها به آسانی قابل انجام است بلکه خروجی بسیار مطلوب‌تری نسبت به روش‌های استفاده از مدل‌های آماری دارد.

۶-۲. پیشنهادات

نکات زیر به عنوان پیشنهاد بیان می‌گردد که این نقطه نظرات، هم از مطالعات صورت گرفته در منابعی که ذکر شده است، برگرفته شده و هم در برخی موارد پیشنهاد شخصی نگارنده این گزارش است:

۱- یکی از مشکلات پیش رو، بحث انواع ابهام در زبان است؛ هرچند که بسیاری از ابهام‌ها شناخته شده هستند، اما فعالیت و پژوهش برای افزایش دقت در تشخیص و تحلیل این ابهام‌ها همچنان ادامه دارد و محققین در پی رفع آن‌ها هستند. پیشنهاد می‌شود در زبان فارسی، مانند بقیه‌ی زبان‌ها، به این مساله بیشتر پرداخته شود. انتظار می‌رود، دانشجویان فارسی زبان، برای حفظ و نگهداری و اعتلای زبان پارسی، تلاش بیشتری در این زمینه داشته باشند.

۲- با اینکه فعالیت‌های زیادی برای تولید گفتار طبیعی (گفتاری که پیوسته باشد و تکیه‌های مناسب در کلمات و حروف داشته باشد) انجام شده است، اما به خاطر محدودیت‌های موجود ناشی از پردازنده‌ها، حافظه‌ها، حالت‌های خاص زبان و ...، هنوز مدلی منطبق بر واقع ارائه نگردیده است که کاملا طبیعی به نظر برسد؛ هرچند که فعالیت همچنان ادامه دارد و اخیرا نتایج خوبی نیز حاصل شده است.

۳- می‌توان در بازسازی گفتار از زبان محاوره استفاده کرد. یعنی همان‌گونه که ما صحبت می‌کنیم، رایانه نیز حرف بزند.

۴- با استفاده از روش پیشنهادی می‌توان یک سیستم متن به گفتار با خروجی بسیار طبیعی و خوب داشت. پس می‌شود دادگان دایفونی برای حالت‌های مختلف گفتاری تولید کرد. همچنین برای تعیین مرز هجایی و دایفونی نیز دادگان دیگری تهیه کرد. ولی این کار نیاز به یک گروه قوی دارد و یک نفر نمی‌تواند پروژه‌ی به این عظمت را تنهایی انجام دهد.

۵- الزام وجود روش‌های ارزیابی سیستم‌های تبدیل متن به گفتار فارسی و تهیه‌ی محتوا، دستورالعمل‌ها و واسط‌های کاربری و نرم افزارهای مورد نیاز، بسیار زیاد می‌باشد. مثلا نرم-افزاری برای زبان فارسی باید وجود داشته باشد که مقایسه‌ی کیفی دو سیگنال را انجام دهد یا اینکه سیگنال مورد نظر را از جهات مختلف امتیازدهی کند.

پیوست‌ها

چالش‌های پردازش متن

۱- مواجهه با چندمعنایی و چندنقشی بودن کلمات: برخی لغات مانند کلمه " شیر" دارای چندین معنی هستند که با توجه به جمله‌ای که در آن واقع می‌شوند، معنی آن‌ها مشخص می‌گردد. بعضی کلمات نیز مانند"در "و" چرا "علاوه بر چندمعنی بودن،دارای چند مقوله‌ی نحوی یا نقش دستوری نیز، هستند. این ویژگی منجر به بالا رفتن سطح ابهام در متن می‌شود.

۲- حذف کلمات و عبارات به قرینه‌ی لفظی یا معنوی: در بسیاری موارد، کلمات یا عباراتی در یک جمله به قرینه‌ی لفظی یا معنوی حذف می‌شود و شنونده باید با تکمیل بخش‌های حذف شده، معنای عبارت را در ذهن خود بازسازی کند. حذف حروف اضافه" در "در جمله "دکتر خانه نیست".

۳- استفاده از افعال مرکب، اصطلاحات و ضرب‌المثل‌ها: در زبان‌های مختلف افعال مرکب، اصطلاحات، ضرب‌المثل‌ها و استعارات، موارد مشکل‌آفرین در پردازش متون هستند چرا که معمولا معنایی کاملاً متفاوت با معنای ظاهریشان دارند. به علاوه اجزای چنین ترکیبی لزوماً در جمله به دنبال هم ظاهر نمی‌شوند و ممکن است بین آن‌ها کلمه یا حتی جمله‌ی دیگری واقع شود و این امر، یافتن و پردازش سازه‌ی مرکب در کل جمله را مشکل می‌کند.

۴- تعیین طبقه‌ی اسامی: گاهی اسامی برخلاف ویژگی‌های ظاهری، طبقه‌ی خود را تغییر می‌دهند. مثلاً در جمله‌ی "خاک‌ها را بِریز توی باغچه " کلمه‌ی خاک‌ها گر چه علامت جمع دارد ولی تعدّد و شمارش را القاء نمی‌کند بلکه بیشتر به مفهوم نکره‌ی جنس و کلیت اشاره دارد. همچنین تشخیص اسامی عام و خاص در کاربردهای مختلف آنها ممکن است ساده نباشد. مانند استفاده از اسامی خاص در نقش عام (ایران سرزمین مولوی‌هاست) و یا اسامی عام در نقش خاص (کلمه‌ی پروانه به عنوان اسم دختران)

۵- بی‌ترتیب‌بودن واحدهای زبان: اگرچه خیلی از زبان‌ها و زبان فارسی دارای ترتیب مرکزی فاعل-مفعول-فعل است ولی دارای استثنائات فراوان و مکرر در ترتیب کلمات می‌باشد. این مسئله باعث می‌شود ساخت دستور زبان مدون و محاسباتی برای زبان و در نتیجه تجزیه و تحلیل نحوی جملات مشکل شود. مثلا جمله‌ی ساده‌ی " دیروز من کتاب را در مدرسه به مریم دادم" می‌تواند به انواع اشکال مختلف با جابجایی متمم‌ها و قید در طول جمله نوشته شود (مانند " من دیروز کتاب را در مدرسه به مریم دادم" یا " من دیروز در مدرسه کتاب را به مریم دادم" یا "من کتاب را در مدرسه دیروز به مریم دادم" ویا "دیروز من در مدرسه کتاب را به مریم دادم".)

۶- کسره‌ی اضافه و حذف آن: یکی از مشکلاتی که بیشتر مربوط به زبان فارسی است این است که در زبان فارسی کسره‌ی اضافه که معمولاً حذف می‌شود دارای چند نقش است . این علامت، صفت و موصوف و یا مضاف و مضاف‌الیه را به هم مرتبط می‌نماید و در انگلیسی «of»و «’s» در بعضی موارد علامات معادل این کسره هستند و در برخی موارد دیگر (مثل اتصال صفت و موصوف) معادلی در انگلیسی ندارد. تشخیص میان نقش‌های مختلف این علامت توسط ماشین کار ساده‌ای نیست، به خصوص زمانی که سامانه و برنامه‌مان دانش معنایی و کاربردی اندکی داشته باشد. در ضمن حذف کسره‌ی اضافه در نوشتار، منجر به ایجاد مشکل در تشخیص مرزهای عبارات اسمی می‌شود.

۷- عدم تطابق اجزای جمله: از لحاظ نظری لازم است میان اجزای مختلف جمله تطابق‌هایی برقرار باشد. مانند مطابقه‌ی فاعل و فعل از جهت تعداد، مطابقه‌ی اجزای جمله و اجزای عبارات اسمی از نظر معنایی. در بعضی موارد برخی از این تطابقات بدون ایجاد خدشه به ساختار معنایی جمله نادیده گرفته می‌شوند. مثل استفاده از فعل جمع برای فاعل مفرد در حالت احترام مانند: آقای مدیر آمدند و یا فعل مفرد برای فاعل جمع غیرجاندار مانند: برگ‌ها می‌ریزد.

۸- وجود ساختار جملات یکسان با معانی و نقش‌های متفاوت: در زبان‌هایی مانند فارسی بسیاری نقش‌های دستوری، با علامت و حرف اضافه مشابه مشخص می‌شوند. برای مثال نقش‌های همراه، حالت، ابزار، وسیله، مقابله یا معاوضه، داشتن و ... با حرف اضافه "با" ظاهر می‌شوند. لذا جملات ذیل اگر چه از لحاظ ظاهری مشابه‌اند، ولی دارای نقش‌های موضوعی متفاوتی هستند و گاهی برای تشخیص این نقش، نیاز به دانش معنایی و کاربردی داریم. برای مثال "با " در جملات" علی با ناراحتی رفت".،" علی با لباس سیاه رفت" ، " علی با حسن رفت ". و " علی با اسبش رفت ".به ترتیب نشان‌دهنده‌ی حالت فعل، توصیف فاعل، همراه و ابزار است.

۹- مشکلات ناشی از ابهام زبان طبیعی: از ویژگی‌های زبان طبیعی وجود ابهام در آن (حتی برای خواننده‌ی انسانی) است که برخی از انواع آن در زیر بر شمرده شده‌اند :

- اول: ابهام ساختاری که معمولا ناشی از ابهام در دسته‌بندی کلمات در عبارات ایجاد می‌شود مانند ابهام در "بطری شیر خشک" یا "ما همه کار می کنیم" (ما همه، کار می‌کنیم یا ما،همه، کار می‌کنیم) ویا "بازجویی او به درازا کشید" ویا " زن و مردِ پیر".
- دوم: ابهام واژگانی ناشی از معانی مختلف یک کلمه مثل ابهام در عبارت "زنِ خیاط".
- سوم: ابهام ارجاعی یا ابهام در مرجع ضمیر مثل ابهام در تشخیص مرجع" او "در جملات " علی حسن را دید .کتاب او روی میز بود" که معلوم نیست ارجاع "او" به علی است یا حسن.
- چهارم: ابهامات ناشی از حذف به قرینه‌ی لفظی یا معنایی. مثلاً در جمله‌ی " در آنجا سعدی شاعر بزرگ ایران را دید" فاعل حذف شده و تشخیص اینکه سعدی فاعل است یا مفعول بدون داشتن دانش پیش‌زمینه در مورد سعدی و موضوع بحث میسر نیست.
- پنجم: ابهام ناشی از صفت جانشین اسم مانند ابهام در عبارت "عصای سخنران" که معلوم نیست که این عصا است که سخنرانی می‌کند و یا عصا متعلق به سخنران است!

- ششم: ابهام ناشی از محدودیت‌های نوشتاری (مانند حذف اِعراب) در کلماتی بروز می‌کند که رسم الخط یکسان ولی تلفظ و معنای متفاوت دارند. این ویژگی در زبان فارسی که حرکات و اعراب حروف در نوشتار حذف می‌شوند، مشهودتر و فراوان‌تر است (مانند کلمات حُکم، حِکَم و حَکَم که دارای نوشتار یکسان و تلفظ و معنای متفاوتند.)

انواع مدل‌های داده‌گرا

۳-۴-۱. مدل‌هایی بر اساس بیشترین مقدار آنتروپی

یکی از پارامترهایی که در بحث‌های آماری مورد استفاده قرار می‌گیرد، آنتروپی است.[۶] در واقع آنتروپی می‌تواند بیانگر نوعی از فشردگی اطلاعات باشد. از آنجا که در بحث‌های پردازشی کامپیوتر، داده‌ها گسسته و اغلب محدود هستند، استفاده از این روش برای تصمیم گیری، بسیار مفید است. برای فهم بهتر آنتروپی، فرض کنید دو تصویر در اختیارمان باشد؛ تصویر اول یک متن تایپ شده با اندازه‌ی مثلا ۶۰۰ در ۶۰۰ پیکسل و تصویر دوم یک منظره با همان اندازه‌ی قبلی باشد. طبق تعریفی که در زیر آمده است، آنتروپی تصویر دوم، که در واقع اجزای بیشتری دارد، بالاتر خواهد بود. یعنی میزان اطلاعاتی که تصویر دوم دارد، نسبت به تصویر اول، به مراتب افزون‌تر است. فرمول آنتروپی به شکل زیر است:

$$H(X) = -\sum_{i} P(x_i) \log P(x_i) \qquad \text{پ-۱}$$

که در رابطه‌ی بالا، حرف p بیانگر مقدار احتمال وقوع یک پیشامد مانند xi است. همچنین این روش به نام‌های *log-linear* و *Gibbs* و *exponential* و *multinomial logit models* نیز شناخته می‌شود. رابطه‌ای به صورت زیر بین آنتروپی و توزیع‌نمایی وجود دارد که از آن، به عنوان یک معیار برای تشخیص استفاده می‌شود:

$$p(x) = \frac{\exp\left(\lambda^T f(x)\right)}{\sum_{y \in X} \exp\left(\lambda^T f(y)\right)}$$ **پ-۲**

که در این رابطه، آرگومان تابع نمایی، ضرب داخلی بین بردار پارامترهای مدل و بردار ویژگی‌های یک مدل است. (برای اطلاع بیشتر به [۶] رجوع شود.)

همچنین در [۱۶] توضیحاتی مفصل، پیرامون: تخمین پارامترها به کمک آنتروپی، استفاده از مدل آنتروپی در دسته‌بندی کردن منابع اطلاعاتی، مدل‌های زنجیره‌ای، مدل‌های تجزیه که در آن از آنتروپی برای جداکردن ویژگی‌های مختلف استفاده شده و ... داده شده است.

از کاربردهای مدل آنتروپی در زبان طبیعی در سال‌های اخیر می‌توان به مواردزیر اشاره نمود [۶]:

- جدا کردن مرز بین جملات[۱].
- استخراج ویژگی‌های آماری گرامر[۲].
- تجزیه‌ی اجزای جمله به قسمت‌های کوچکتر[۳].
- برچسب‌گذاری قسمت‌های صوتی[۴].
- و ... .

۳-۴-۲) درخت‌های تصمیم‌گیری

درخت‌های تصمیم‌گیری[۵] ابزاری هستند برای جداکردن قسمت‌های مرتبط به یک مفهوم به‌طوری که آن‌را به قسمت‌های مختلف و زیرمجموعه‌هایی از آن تقسیم می‌کنند [۶و۸] به عنوان مثال می‌توان

---

[1] *Sentence boundary detection*

[2] *stochastic attribute value grammars*

[3] *parse selection*

[4] *part-of-speech tagging*

[5] *Decision Trees*

به نمودارهای درختی که برای اجزای یک جمله در نظر گرفته می‌شود اشاره کرد. البته در اینجا؛ منظور از نمودارهای درختی، استفاده از آنها به عنوان ابزاری برای تصمیم‌گیری است.

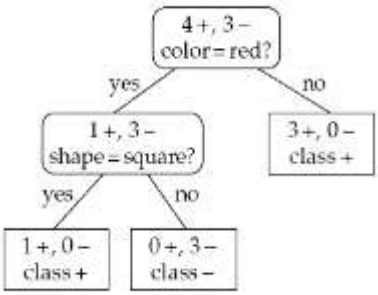
به عنوان مثال، ابتدا جدولی در ادامه ارائه شده؛ و سپس یک نمودار درختی متناظر با آن برای تصمیم‌گیری ترسیم شده‌است.

توجه به این دو نکته نیز مهم است که در استفاده از درخت‌های تصمیم‌گیری، از دو مفهوم: آموزش داده‌ها[1] (برای غنی‌کردن پایگاه داده‌ها) ، و آنتروپی (معیاری برای تصمیم‌گیری) به وفور استفاده می‌شود.

Size	*Color*	*Shape*	*Class*
medium	blue	circle	+
small	red	square	+
large	green	trapezoid	+
large	green	square	+
small	red	triangle	−
large	red	triangle	−
large	red	trapezoid	−

شکل ۵ : نمونه‌ای از یک سری داده برای استفاده در درخت تصمیم‌گیری [۶]

یک درخت نوعی برای این جدول می‌تواند به صورت زیر ارائه شود :



شکل۶ : مدل یک درخت تصمیم‌گیری؛ درخت تصمیم‌گیری برای تعیین اینکه با توجه به داده‌ی ورودی، و جدول، آیا رنگ قرمز درست تشخیص داده شده است؟ [۶]

یکی از مفاهیمی که در این‌جا استفاده می‌شود، هرس‌کردن[1] درخت است. یعنی کم‌کردن شاخ و برگ‌های اضافی از درخت که سبب افزایش سرعت در تصمیم‌گیری‌های پیاپی خواهد شد[۶] .

[1] *Training*

موارد استفاده از درخت تصمیم‌گیری در پردازش زبان طبیعی که در [۶و۸] بـه آن اشـاره شـده اسـت؛ عبارت‌است از:

- تبدیل یک حرف نوشتاری به واج (گفتاری) متناظر.
- برچسب‌گذاری مقوله واژگانی[۲] .
- نشان‌گذار متن[۳].
- برچسب‌گذاری متن[۴].
- مدل‌کردن زبانی[۵].
- تخمین‌گرهای جداکننده‌ی عبارات (بیشتر در حوزه‌ی متن.)
- تشخیص هرزنامه[۶].

۳-۴-۳) شبکه‌های عصبی مصنوعی

از شبکه‌های عصبی مصنوعی[۷] ، در پردازش زبان طبیعی، به اشکال مختلفی استفاده می‌شود. البته در این‌جا از مدل‌های آماری آن، بیشتر استفاده می‌شود. یعنی شبکه‌های عصبی مصنوعی، بیشتر در قالب مدل‌های آماری استفاده می‌شوند. یکی از موفق‌ترین کاربردهایی که در پردازش زبان طبیعی داشته است، مدل‌کردن زبان طبیعی و تجزیه‌ی زبان طبیعی به اجزای کوچکتر و زیرمجموعه‌های دسته‌های بزرگتر بوده است. [۶]

---

[1] *Recision*

[2] *part of speech tagging*

[3] *Tokenization*

[4] *Parsing*

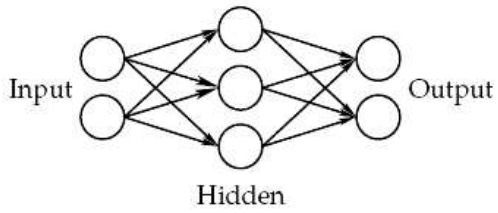
[5] *language modeling*

[6] *Spam*

[7] *Artificial Neural Networks*

عبارت شبکه‌ی عصبی مصنوعی به حالتی اطلاق می‌شود که در آن، انواع مختلفی از مدل‌های محاسباتی، ویژگی‌های مشخصی را، با الهام از نورون‌ها در مغز انسان، با همدیگر به اشتراک می‌گذارند. در واقع واحدهای محاسباتی ساده‌تر به کمک نورون‌ها (به عنوان واسط) به همدیگر مربوط شده و واحدهای محاسباتی با توانایی‌های بالاتری را می‌سازند. البته لازم به ذکر است که برای کارآمدترشدن مدل‌ها، مدل‌های طراحی شده، باید توسط داده‌ها آموزش داده شوند تا به بیشترین میزان کارایی برسند.

یکی از مفاهیمی که در این شاخه از علوم استفاده می‌شود، ادراک چند لایه‌ای[1] است. این مدل چهار پارامتر مختلف دارد: ورودی، خروجی، واسط‌های وزن‌دار، لایه‌ی پنهان. که به صورت یک مسیر جلوسو[2] با هم‌دیگر در ارتباط هستند. شکل زیر این رابطه ها را نشان می‌دهد.



شکل۷ : مدل *MLP* در شبکه های عصبی

اگر مقداری که بر روی هر نقطه (گره) قرار دارد را j بنامیم و وزن‌دار ورودی به آن از گره‌ی i را Wij بنامیم، مقدار Zj که متناظر با خروجی بر روی گره j است از رابطه زیر بدست می‌آید:

$$z_j = \frac{1}{1 + \exp\left(-\left(w_{j0} + \sum_i z_i w_{ji}\right)\right)}$$

برای اطلاعات بیشتر از شبکه‌های عصبی و کاربرد آن در پردازش زبان طبیعی به [۷و۸] مراجعه شود. در ادامه برخی از کاربردهای شبکه عصبی در پردازش آورده شده‌است:

[1] *Multi-layered perceptron*

[2] *Feed Forward*

- مدل‌سازی زبان[1]
  - مدل‌های شبکه عصبی زبان[2].
  - مدل‌های ترکیبی (نحوی) عصبی زبان[3].
- تجزیه[4] به اجزای کوچکتر
  - تجزیه‌ی انتخابی[5].
  - تجزیه‌ی وابسته [6].
  - تجزیه‌ی عمل‌گرایانه و نقش‌های معنایی[7].
  - برچسب‌گذاری نقش‌های معنایی[8].

۳-۴-۴) مدل مارس[9]

روشی برای تخمین یک تابع با بعد بالا، با داده‌های خلوت (پراکنده) است که از روی داده‌ها، پارامترها و ساختار مدل محاسبه می‌شود و تعامل بین متغیرها، به‌خوبی مشخص می‌گردد.

این روش علاوه بر تخمین خوبی که دارد، می تواند تفسیر خوبی نیز از مدل ارائه دهد. بدین صورت که یک رابطه‌ی ریاضی برای تخمین تابع تولید می‌کند که در این معادله دو مفهوم اساسی تابع پایه و حداکثر درجه تعامل، وجود دارد.

---

[1] *language modeling*

[2] *Neural network language models*

[3] *Neural syntactic language models*

[4] *Parsing*

[5] *Constituency parsing*

[6] *Dependency parsing*

[7] *Functional and semantic role parsing*

[8] *Semantic role tagging*

[9] *Mars*

این مدل برای مدل‌سازی مقدار زمان کشش واج و مدل‌سازی منحنی گام استفاده می‌شود. برای اطلاعات بیشتر به [۸] مراجعه شود.

۳-۴-۵) ماشین بردار پشتیبان

این روش برای اولین بار در مسائل طبقه‌بندی، مورد استفاده قرار گرفت. از ویژگی‌های این روش؛ قدرت تعمیم بالا و کارایی خیلی‌خوب آن در مسئله‌های طبقه‌بندی و تقریب است. از این رو، در بحث سنتز گفتار، در مواردی همچون مدل‌سازی دیرش زمانی واج، از آن استفاده می‌شود.

این روش با نگاشت غیر خطی داده‌ها به یک فضای با ابعاد بالا، عمل تقریب را انجام می‌دهد و در آن، توابع به وسیله‌ی یک‌سری از داده‌هایی که متعلق به مجموعه‌ی یادگیری هستند تخمین زده می‌شوند که به آن داده‌ها، بردار پشتیبان گفته می‌شود. تعداد این بردارها به دقت زیاد، بستگی دارد. برای اطلاعات بیشتر به [۶] مراجعه شود.

۳-۴-۶) مدل مخفی مارکوف

ایده‌ی اصلی مدل مخفی مارکوف در دهه‌ی ۶۰ میلادی مطرح شد؛ و به خوبی در مسائل مختلف بازشناسی گفتار، به کار گرفته شده است. مدل مخفی مارکوف برای مدل‌سازی دنباله‌های تصادفی متناهی به‌خوبی مورد استفاده قرار می‌گیرد [۸].

فرض اساسی استفاده از این مدل، در نظر گرفتن دنباله‌ی ورودی به صورت یک فرآیند تصادفی است. از آن‌جا که این فرآیند مشاهده‌پذیر نیست، از طریق دسته‌ی دیگری از فرآیندهای تصادفی مرتبط با آن فرآیند، که دنباله‌ی مشاهدات را تولید می‌کنند؛ مورد بررسی قرار می‌گیرد.

مدل مخفی مارکوف به عنوان یکی از موفق‌ترین روش‌های آماری، خصوصاً در بازشناسی گفتار، شناخته شده است. توانایی مدل‌های مخفی مارکوف در مدل‌سازی تغییرات زمانی سیگنال گفتار و پشتوانه ریاضی این مدل‌ها، آن‌ها را به عنوان انتخابی بی‌رقیب در مدل‌سازی آکوستیکی مطرح می‌کند.

مراجع

[01] Sohrab Hojjatkhah, Ali Jowharpour, "Segmentation Words for Speech Synthesis in Persian Language Based On Silence," *International Journal of Recent Research in Mathematics Computer Science and Information Technology*,Vol. 1, Issue 1, pp. 14-17, April - June 2014.

[۰۲] شورای عالی اطلاع‌رسانی، *بررسی ابعاد و تفاوت‌های پیکره‌های برچسب داده‌ای و پیکره‌های خام در زبان فارسی، فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی*، انتشارات دانشگاه علم و صنعت ایران، تهران، ایران، ۱۳۸۸.

[۰۳] شورای عالی اطلاع‌رسانی، " امکان‌سنجی پروژه‌های زیرساختی کاربری خط و زبان فارسی در محیط رایانه‌ای،" *فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی*، دانشگاه علم و صنعت ایران، ۱۳۸۸.

[۰۴] مهدی محسنی، بهروز مینایی بیگدلی، "سیستم برچسب‌گذاری اجزای واژگانی کلام در زبان فارسی،" *دوفصل‌نامه‌ی پردازش علائم و داده ها*، شماره‌ی دوم، پیاپی دوازدهم، صص. ۱۳-۲۶، ۱۳۸۸.

[۰۵] شورای پژوهش‌های علمی کشور، *گزارش نهایی طرح ملی پردازش زبان فارسی*، انتشارات دانشگاه صنعتی امیرکبیر، تهران، ایران، بهار ۱۳۸۱.

[۰٦] محمدمهدی همایون‌پور، *پژوهشنامه تبدیل متن به گفتار*، شورای عالی اطلاع‌رسانی، تهران، ایران، ۱۳۹۰.

[07] Matthews, Clive, *An Introduction to Natural Language Processing Through Prolog Learning About Language*, Addison Wesley Publishing Company, London,England, 1998.

[08] Paul Taylor, *Text to Speech Synthesis*, University of Cambridge, 2007.

[09] Ehsan Rasekh, Mohammad Eshghi, "A Syllable-Phoneme Solution for Efficient Farsi Text-to-Speech Systems," *Proceeding of the 10th IEEE Conference on TENCON*, pp. 171-175, Hong Kong, November 2006.

[10] Ehsan Rasekh, Mohammad Eshghi, "An Efficient Network for Farsi Text to Speech Conversion Using Vowel State," *Proceeding of the 10th IEEE Conference on TENCON*, pp. 176-179, Hong Kong, November 2006.

[11] Mohammad Mehdi Arab, Ali Azimizadeh, “Construction of a Persian Letter-To-Sound Conversion System Based on Classification and Regression Tree in festival,” *proceedind of Computational Approaches to Arabic Script Languages (CAASL-3*), pp.132-138, Stanford University, 2009.

[12] Iman Rasekh, Ehsan Rasekh, Mohammad Eshghi, "An Approach to Recognize and Pronounce Words with Alternative Pronunciations in Farsi," *Proceeding of the 23rd Canadian Conference on Electrical and Computer Engineering (CCECE*), pp.1-4, Calgary, May 2010.

[۱۳] محمدمهدی همایون‌پور، *ارائه‌ی داده‌ها، دستورالعمل و نرم افزارهای ارزیابی عملکرد سیستم‌های تبدیل متن به گفتار فارسی*، انتشارات دانشگاه صنعتی امیرکبیر، تهران، ایران، ۱۳۸۹.

[۱۴] وحید صادقی، "طراحی و ارزیابی یک سامانه بازسازی گفتار به روش هم‌گذاری واحدهای حساس به بافت نوایی،" *مجله پردازش علائم و داده‌ها*، شماره‌ی دوم، پیاپی چهاردهم، صص. ۶۹-۸۴، ۱۳۸۹.

[۱۵] محمد مهدی همایون پور، مجید نم نبات، " فارس بیان: گامی اساسی در سنتز گفتار فارسی،" *دومین کارگاه پژوهشی زبان و رایانه*، دانشگاه صنعتی امیرکبیر، تهران، ایران، تیرماه ۱۳۸۵.

[۱۶] شهرام کلانتری، محمد مهدی همایون پور، "افزایش سرعت سنتز گفتار در روش انتخاب واحد با استفاده از جستجوی شعاعی و پیرایش دادگان،" *مجله پردازش هوشمند علائم*، شماره‌ی دوم، پیاپی چهاردهم، صص. ۴۹-۵۸، آبان ۱۳۸۹.

[17] C. C. Wai, *Speech Coding Algorithms*, John Wiley & Sons, Inc, 2003.

[۱۸] تقی وحیدیان کامیار، *نوای گفتار (تکیه، آهنگ، مکث) در فارسی*، انتشارات دانشگاه فردوسی مشهد، ایران، مشهد، زمستان ۱۳۷۹.

[۱۹] منصور شیخان، مجید نصیرزاده، علی دفتریان، "مراحل و نحوه تهیه دادگان‌های صوتی هجایی و دایفونی برای سامانه تبدیل متن به گفتار فارسی،" *نشریه دانشکده مهندسی امیرکبیر*، سال هفدهم، شماره‌ی دوم، صص. ۳-۱۲، ۱۳۸۴.

[۲۰] غزال شیخی، سید حمید محمودیان، "تقطیع هجایی گفتار پیوسته فارسی با استفاده از آستانه گذاری ضرایب موجک و نرم سازی فازیِ تابع انرژی،" *نشریه روش‌های هوشمند در صنعت برق*، سال چهارم، پیاپی پانزدهم، صص. ۲۰-۳۲، پاییز ۱۳۹۲.

[۲۱] سید سعید آیت، "طراحی و پیاده سازی دادگان دایفون زبان فارسی برای کاربرد زبانشناسی رایانه‌ای،" *مجله پژوهش‌های زبان‌شناسی*، سال دوم، شماره‌ی دوم، صص. ۱-۱۱، پاییز و زمستان ۱۳۸۹.

[۲۲] منصور شیخان، مجید نصیرزاده، علی دفتریان، "طراحی و پیاده‌سازی سیستم تبدیل متن به گفتار طبیعی برای زبان فارسی،" *نشریه دانشکده مهندسی*، شماره هفدهم، شماره‌ی دوم، صص. ۳۱-۴۸، ۱۳۸۴.

ABSTRACT

In this thesis we propose a new speech synthesis method based on diphones. Our motivation for focusing on diphones is their desired characteristics including their limited numbers and ease and clarity of transition.

We aim to produce different types of sentences at the output of the speech synthesizer. Most of existing approaches on Farsi language, consider complicated approaches such as Hidden Markov Model (HMM), decision trees, and Neural Networks (NN). We aim to show that using diphones we will be able to produce different types of sentences with more naturalness. We consider seven different types of diphones with unique characteristics.

In our simulations, we first combined diphones extracted from recorded sentences to generate the target sentences. Then we played some sample audios for some audiences and asked them to assign a score, regarding to naturalness and clarity, to compute mean opinion score (MOS) measure.

Comparison of our results with that of some conventional approaches, some were based on decision tree and triphones, shows that our proposed method received a higher score on clarity. We also note that there are possible approaches to improve the naturalness of generated sentences. Details on such approaches are discussed in this thesis.

Keywords: Farsi text to speech conversion, speech synthesis, diphone, naturalness, MOS score

University of Shahrood

Faculty of Robotic and Electrical Engineering

Diphone extraction of various modes of speech to improve speech system synthesis

Mohammadreza Haghighat

Supervisor:

Dr. Hadi Grailu

October 2015