بسم الله الرحمن الرحيم

دانشگاه صنعتی امیر کبیر
پلی تکنیک تهران
دانشکده مهندسی کامپیوتر و تکنولوژی اطلاعات

پایان نامه کارشناسی ارشد مهندسی کامپیوتر
گرایش هوش مصنوعی

مدل کردن نوای گفتار فارسی با استفاده از روشهای داده گرا و قانونگرا

نگارش:
معصومه بحرینی

استاد راهنما: دکتر محمد مهدی همایون پور

آبان ۱۳۸۶

بسمـه تعالـــی

دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)
معاونت پژوهشی

فرم اطلاعات پایان نامه
کارشناسی ارشد و دکترا

تاریخ: .....

پیوست: .....

نام و نام خانوادگی: معصومه بحرینی	**دانشجوی آزاد** ☐	**بورسیه** ☐	**معادل** ☐	
شماره دانشجوئی: ۸۳۱۳۱۲۰٤	**دانشکده:** مهندسی کامپیوتر	**رشته تحصیلی:** هوش مصنوعی		
نام و نام خانوادگی استاد راهنما: دکتر محمد مهدی همایونپور				
عنوان پایان نامه به فارسی: مدل کردن نوای گفتار فارسی با استفاده از روشهای داده گرا و قانون گرا				
عنوان پایان نامه به انگلیسی: Farsi prosody modeling using data driven and rule based methods				
نوع پروژه: کارشناسی ارشد ☐ دکترا ☐	**کاربردی** ☐	**بنیادی** ☐	**توسعه ای** ☐	**نظری** ☐
تاریخ شروع: بهمن ۱۳۸٤	**تاریخ خاتمه:** شهریور ۱۳۸۶	**تعداد واحد:** ۶		
سازمان تأمین کننده اعتبار:				
واژه های کلیدی به فارسی: نوای گفتار، کشش، انرژی، فرکانس پایه، ماشین پشتیبان بردار، مارس				
واژه های کلیدی به انگلیسی: prosody, duration, intensity, fundamental frequency, SVM, MARS				

نظرها و پیشنهادها به منظور بهبود فعالیت های پژوهشی دانشگاه:

استاد راهنما:

دانشجو:

امضاء استاد راهنما: **تاریخ:**

نسخه ۱: معاونت پژوهشی
نسخه ۲: کتابخانه و به انضمام دو جلد پایان نامه به منظور تسویه حساب با کتابخانه و مرکز اسناد و مدارک علمی

تقدیم به

همسرم، دخترم و پدر و مادرم

تقدیر و تشکر

اکنون که به یاری خداوند متعال و با کمک ها و مساعدت های بی دریغ استاد محترم، دکتر محمدمهدی همایون‌پور، این پایان‌نامه به اتمام رسیده است، بر خود واجب می‌دانم از زحمات ایشان، که با راهنمایی‌ها و کمک‌هایشان بنده را در انجام این کار، یاری کردند کمال تشکر و امتنان را داشته و برای ایشان آرزوی موفقیت و سلامت دارم.

همچنین از آقای سینا ایران‌نژاد به خاطر ارائه نقطه نظرهای مفیدشان در زمینه بخش‌های مختلف انجام این مهم، کمال تشکر و سپاس را دارم.

متواضعانه ارادت و سپاس خود را از تمامی اساتیدی که در طول دوران تحصیل، به خصوص در مقطع کارشناسی ارشد، از محضرشان کسب فیض نموده‌ام، عرضه می‌دارم.

چکیده

هدف از انجام این پایان‌نامه، مدلسازی نوای گفتار فارسی با استفاده از روش‌های داده‌گرا، برای سیستم‌های تبدیل متن به گفتار فارسی می‌باشد. روش‌های داده‌گرای بکار گرفته شده، شامل منحنی‌های متعدد تقریب انطباقی (مارس)، شبکه عصبی و ماشین پشتیبان بردار می‌باشند. مارس، تکنیکی برای تخمین یک تابع با بعد بالا با داده‌های خلوت می‌باشد که از روی داده‌ها پارامترها و ساختار مدل را بدست می‌آورد و قابلیت تفسیر مدل را فراهم می‌کند. ماشین پشتیبان بردار قابلیت تعمیم بسیار بالایی دارد به‌طوری‌که در اکثر موارد، کارایی آن در آموزش و تست، تقریبا یکسان می‌باشد. شبکه عصبی در محیط‌های نویزی خیلی خوب عمل می‌کند اما امکان تفسیر خروجی ندارد.

نوای گفتار شامل دیرش، فرکانس پایه و انرژی آن می‌باشد که معمولا مقدار دیرش برای هر واج گفتار تخمین زده می‌شود و فرکانس پایه و انرژی به صورت یک منحنی برای کل گفتار، تولید می‌شود.

مقدار دیرش هر واج، با استفاده از روش مارس، شبکه عصبی و ماشین پشتیبان بردار تخمین زده شد و با استفاده از نتایج مارس، اهمیت عوامل موثر در کشش و تعامل بین عوامل، مورد تحلیل واقع شد. با توجه به زیاد بودن تعداد داده‌ها و سرعت پایین ماشین پشتیبان بردار در آموزش و آزمایش، دو شیوه متفاوت بکار گرفته شد. در روش اول با استفاده از چندی‌سازی برداری در فضای ورودی، تعداد داده‌های آموزشی به میزان قابل توجهی کاهش یافت و در روش دوم، فضای خروجی با توجه به مقدار دیرش هر داده، به چند خوشه تقسیم شد و برای هر خوشه، یک مدل تخمین جداگانه، ایجاد گردید. هر دو روش زمان آموزش و تست سیستم را با حفظ کارایی کاهش دادند.

به منظور تولید منحنی گام، از روش فوجی‌ساکی، تیلت و منحنی‌های قطعه‌قطعه استفاده شد. روش فوجی‌ساکی برای منحنی گام، دو جزء دستورات تکیه و عبارت را فرض می‌کند که هر کدام دارای پارامترهای خاص خود هستند. پارامترهای دستورات تکیه، برای هجاهای تکیه‌بر و پارامترهای دستورات عبارت، برای اولین هجای عبارت‌های نوایی گفتار تخمین زده می‌شود و با استفاده از این پارامترها، منحنی گام با بکارگیری فرمول فوجی‌ساکی، تولید می‌شود. به منظور تخمین پارامترها، روش‌های مارس، شبکه عصبی و ماشین پشتیبان بردار بکار گرفته شدند که نتایج آزمایش‌ها نشان داد، روش مارس قادر به تخمین کلیه پارامترهای فوجی‌ساکی نمی‌باشد.

مدل تیلت، منحنی گام را به صورت دنباله‌ای از رویدادهای آهنگین فرض می‌کند. رویدادهای اصلی شامل تکیه زیروبمی (a) و نواخت‌های مرزی (b) هستند. هر دو نوع رویداد با پارامترهای زمان شروع رویداد، فرکانس پایه در لحظه شروع رویداد، میزان دیرش، اندازه و عدد تیلت مدل می‌شوند. با استفاده از این پارامترها و یکسری فرمول، شکل کنتور F0 برای این رویدادها تولید می‌شود و سپس با اتصال کلیه رویدادها به یکدیگر، کل منحنی گام تولید می‌شود. با استفاده از روش‌های یادگیری ماشین، پارامترهای تیلت برای کلیه هجاهای متن تخمین زده شدند.

در روش منحنی‌های قطعه‌قطعه، برای هر واج منحنی گام تولید می‌شود و از اتصال کلیه این منحنی‌ها، منحنی گام برای کل گفتار بدست می‌آید. منحنی هر واج با استفاده از چند جمله‌ای درجه دوم تولید می‌شود و برای تخمین ضرائب این چند جمله‌ای‌ها، از روش‌های داده‌گرا استفاده می‌شود.

در زمینه انرژی گفتار، ابتدا عوامل تاثیرگذار روی مقدار انرژی بررسی گردید و سپس با استفاده از آن عوامل، به مدلسازی منحنی انرژی پرداخته شد. منحنی انرژی گفتار نیز، با استفاده از روش منحنی‌های قطعه‌قطعه مدلسازی گردید که در آن برای هر واج، منحنی انرژی‌اش تولید می‌شود و از اتصال این منحنی‌ها، منحنی انرژی کل گفتار بدست می‌آید. منحنی هر واج با استفاده از چند جمله‌ای درجه دوم تولید می‌شود و برای تخمین ضرائب این چند جمله‌ای‌ها، از روش‌های داده‌گرا استفاده می‌شود.

به منظور ارزیابی نتایج، تست شنیداری MOS و همچنین معیارهای ضریب همبستگی و میانگین مربع خطا، محاسبه شد.

فهرست مطالب

فهرست شکل‌ها

فهرست جدول‌ها

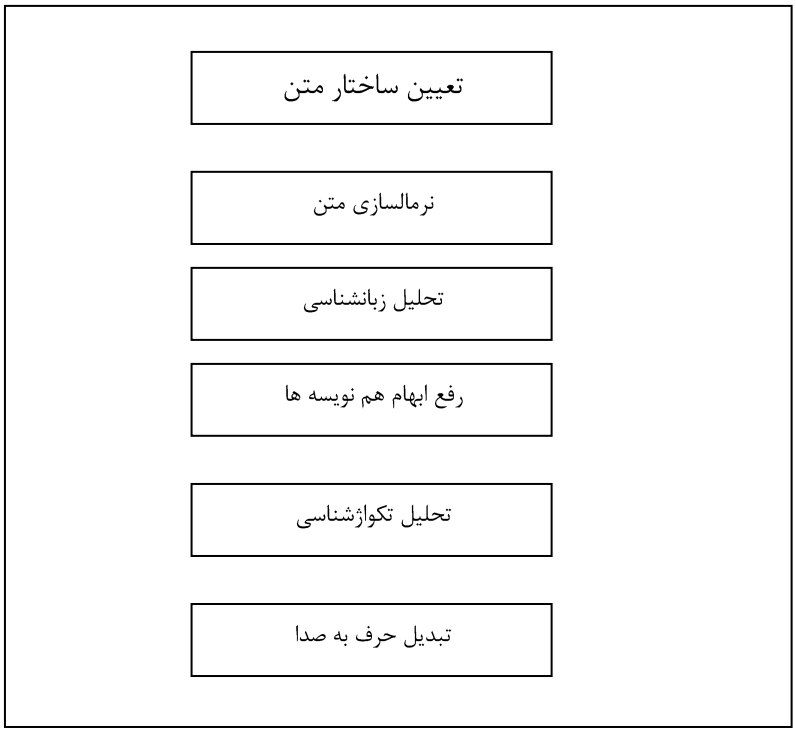
فصل اول: سیستم تبدیل متن به گفتار و اجزاء آن

۱ -سیستم تبدیل متن به گفتار و اجزاء آن

مبدل متن مبدل متن به گفتار، یک سیستم کامپیوتری می‌باشد که باید قادر به خواندن هر گونه متنی که به آن داده می‌شود، باشد. یک سیستم تبدیل متن به گفتار که به اختصار TTS[1] نامیده می‌شود، شامل سه بخش می‌باشد: تحلیلگر متن، تولید کننده نوا و سنتز کننده گفتار. جزء تحلیلگر متن، متن ورودی را از نظر نحوی یا مفهومی بررسی می‌کند و یکسری ویژگی‌های زبانشناسی از آن استخراج می‌کند، این عمل را با استفاده از پردازش زبان طبیعی[2] و با انجام تحلیل‌های لغوی، صرفی، نحوی و معنایی انجام می‌دهد. جزء تولید کننده نوا ویژگی‌های زبانشناسی را دریافت کرده و اطلاعات نوایی را تولید می‌کند. این اطلاعات شامل کنتور پیچ، کنتور انرژی، دیرش و مکث می‌باشد. طبیعی بودن یک گفتار سنتز شده بستگی زیادی به اطلاعات نوا دارد. سنتز کننده گفتار اطلاعات نوا را با واحدهای سنتز انطباق می‌دهد، سپس الگوریتم‌های بهبود نوا روی واحدهای سنتز شده اعمال می‌شود تا صدای طبیعی تولید شود. در قسمت‌های بعدی این فصل، هرکدام از بخش‌های یک سیستم تبدیل متن به گفتار، شرح داده می‌شود.

۱ ۱ -تحلیلگر متن

به منظور تبدیل متن به دنباله آوایی مورد نیاز برای بخش سنتز در یک سیستم TTS، تحلیل متنی مورد نیاز است. شکل (۱-۱) دیاگرام یک تحلیلگر متن را در حالت کلی نشان می‌دهد.



شکل۱-۱- دیاگرام یک تحلیلگر متن

در طی فرایند تعیین ساختار متن، ساختارهای درون متن مانند پاراگراف‌ها و جملات و نیز ساختارهای پیشرفته مانند پست الکترونیکی و صفحات وب تشخیص داده می‌شود. نرمالسازی متن عبارتست از فرایندی که در طی آن، متن حاوی واژه‌ها، اعداد، کوته‌نوشته‌ها و... به متنی تبدیل می‌شود که در آن تمام عناصر غیر املایی مورد اشاره به معادل لغوی و بدون ابهامشان تبدیل شده باشند. تحلیل زبانشانسی تجزیه صرفی، نحوی و معنایی متن را انجام می‌دهد. کلماتی که دارای شکل نوشتاری یکسان و تلفظ متفاوت هستند در مرحله رفع ابهام از هم‌نویسه‌ها مد نظر قرار می‌گیرند. فرایند تجزیه کلمه به تکواژهایش

[1] Text-To-Speech
[2] Natural Language Processing- NLP

تحلیل تکواژشناسی نامیده می‌شود که در آن مرحله برای هر کلمه با توجه به تکواژهایش، تلفظ متناظر آن تولید می‌شود. تبدیل حرف به صدا آخرین مرحله در تحلیل متن می‌باشد که در مواردی که دستیابی به معادل آوایی کلمه با هیچیک از تحلیل‌های صرفی، نحوی و معنایی و استفاده از واژگان میسر نباشد، برای هر کلمه معادل آوایی آن را تولید می‌کند. بعضی از بخش‌های تحلیلگر متن در زیر شرح داده شده‌اند.

۱ ۱ ۱ -نرمال سازی متن

یک متن معمولی ممکن است شامل عدد، علائم اختصاری، سمبل‌ها و علائم ویژه باشد که باید نمایش آوایی آنها را در نظر بگیریم. نرمال‌سازی متن عبارتست از فرایندی که در طی آن، متن حاوی واژه‌ها، اعداد و ... به متنی تبدیل می‌شود که درآن تمام عناصر غیر املایی به معادل لغوی‌شان تبدیل شده باشند. در مثال‌های زیر، متن‌های (الف) متن‌های غیر نرمال و متن‌های (ب)، صورت نرمال شده آن می‌باشد.

الف- شماره تلفن منزل وی، ۷۷۵۸۶۸۶ می‌باشد.

ب- شماره تلفن منزل وی، هفتصدوهفتادوپنج، هشتادوشش، هشتادوشش، می‌باشد.

الف- ۷٪ پولش را خرج کرد.

ب- هفت درصد پولش را خرج کرد.

الف- ۲/۳ مالش را بخشید.

ب- دو سوم مالش را بخشید.

مشاهده می‌شود که در نرمال‌سازی، قوانینی که انسان‌ها در هنگام صحبت کردن رعایت می‌کنند را باید مد نظر قرار داد، مثلا افراد معمولا اعداد طولانی مانند شماره تلفن را به صورت شکسته بیان می‌کنند. علائم ریاضی مانند علامت درصد و کسر باید به معادل تلفظی‌شان، تبدیل شوند. همچنین اگر در متن علامت اختصاری وجود داشته باشد مانند (ه.ق.) باید معادل تلفظی آن (هجری قمری) را در نظر گرفت. مرحله نرمال‌سازی متن ورودی را می‌توان با استفاده از یکسری قوانین از پیش نوشته شده، انجام داد.

۱ ۱ ۲ -رفع ابهام از کلمات هم‌شکل (هم نویسه ها)

یکی از مسائلی که درتحلیل متن، زیاد با آن برخورد می‌شود کلمات و اعدادی هستند که تلفظ آنها بدون استفاده از محتوای متنی که کلمه در آن قرار دارد امکان‌پذیر نیست. این کلمات که دارای شکل نوشتاری یکسان و تلفظ متفاوت هستند هم‌نویسه نامیده می‌شوند. مثلا کلمه "ببر" دارای دو تلفظ متفاوت 'babr' و 'bebar' می‌باشد که تلفظ درست این کلمه، با استفاده از متنی که شامل این کلمه می‌باشد مشخص می‌شود. بسیاری از هم نویسه‌ها با استفاده از اطلاعات نوح نحوی (POS) شان می‌توانند رفع ابهام شوند مثلا کلمه ببر اگر درجایگاه اسم باشد 'babr' و اگر در جایاه فعل باشد 'bebar' تلفظ می‌شود. اما مواردی نیز وجود دارد که این قانون نمی‌تواند مشکل را حل نماید مثلا دو کلمه ممکن است دارای نوح نحوی یکسان باشند که در این حالت با استفاده از اطلاعات معنایی متن، می‌توان رفع ابهام نمود. مثلا همان کلمه ببر اگر در جایگاه فعل باشد دارای دو تلفظ متفاوت 'bebar' و 'bebor' می‌باشد که با توجه به اینکه هر دو فعل متعدی هستند نمی‌توان از روی نوح نحوی کلمات قبل و بعدشان رفع ابهام نمود و باید حتما از اطلاعات معنایی استفاده نمود. برای رفع ابهام معمولا از اطلاعات نوح نحوی خود کلمه و همسایگانش، ساختار جمله و... استفاده می‌کنند، حتی ممکن است به اطلاعات پاراگرافی که جمله درون آن قرار گرفته باشد نیز نیاز داشته باشیم. معمولا برای حل اتوماتیک مسئله از روش لیست‌های تصمیم‌گیری [Yarowsky 1997] ، استفاده می‌شود.

۱ ۱ ۳ -تحلیل تکواژشناسی

معمولا در فرهنگ لغت، شکل ساده یک کلمه وجود دارد و صورت‌های صرفی آن و مشتق‌های آن قرار داده نمی‌شود. برای تولید تلفظ کلمه‌ای که شکل اصلی‌اش در فرهنگ لغت وجود دارد اما صورت صرفی‌اش وجود ندارد از تحلیل تکواژشناسی

استفاده می‌شود. تکواژ عبارتست از کوچکترین عنصر معنادار تشکیل دهنده هر کلمه (مانند پسوندها، پیشوندها و ریشه‌ها). در این مرحله کلمه ابتدا به اجزاء تشکیل دهنده‌اش تجزیه می‌شود و سپس برای هر جزء تلفظ مربوطه‌اش تولید می‌شود. فرایند تجزیه کلمه به تکواژهایش، تحلیل تکواژشناسی نامیده می‌شود. مثلا کلمه "شهرت" بدون استفاده از تحلیل تکواژشناسی داری تلفظ 'Sohrat' می‌باشد که اگر تحلیل تکواژشناسی را انجام دهیم تلفظ 'Sahrat' به معنای "شهرِ تو" نیز به تلفظ آن اضافه می‌شود.

۱ ۱ ۴ -تبدیل حرف به صدا

مشخص کردن تلفظ یک کلمه، معمولا تبدیل متن به واج[3] یا املا به واج[4]، نامیده می‌شود و فرایندی است که درآن طرز تلفظ کلمات متن، مشخص می‌شود. سیستم‌های سنتز گفتار از دو روش اصلی برای این هدف استفاده می‌کنند. ساده‌ترین روش، روش مبتنی بر فرهنگ لغت می‌باشد که یک فرهنگ لغت بزرگ شامل تمام کلمات زبان و تلفظ صحیحشان را تهیه می‌کنند و برای تعیین تلفظ هر کلمه به آن مراجعه می‌کنند. روش دیگر، که قانون‌گرا[5] می‌باشد تلفظ کلمات را با توجه به شیوه دیکته-شان و با استفاده از یکسری قانون، مشخص می‌کند. هردو روش دارای امتیازها و معایبی هستند، روش مبتنی بر فرهنگ لغت، سریع و دقیق می‌باشد اما در صورت برخورد با یک کلمه ناشناخته، قابل استفاده نمی‌باشد. همچنین با افزایش اندازه فرهنگ لغت، میزان حافظه مورد نیاز سنتزکننده گفتار، زیاد می‌شود. از طرف دیگر در سیستم قانون‌گرا اگر بخواهیم همه استثناها و موارد خاص را مد نظر قرار دهیم، تعداد قوانین خیلی زیاد می‌شود و پیچیدگی سیستم افزایش پیدا می‌کند.

بعضی زبان‌ها مانند اسپانیایی در نوشتار قوانین منظمی را استفاده می‌کنند بنابراین تولید تلفظ هر کلمه در این زبان‌ها بسیار ساده می‌باشد. معمولا از روش‌های قانون‌گرا برای تولید تلفظ استفاده می‌شود و برای تلفظ اسامی اشخاص، از فرهنگ لغت استفاده می‌کنند. زبان‌هایی مانند فارسی و انگلیسی شیوه نگارش نامنظمی دارند و همیشه آنچه که نوشته می‌شود به همان صورت خوانده نمی‌شود مثلا در کلمه "خواهر" حرف "و" ادا نمی‌شود و یا در کلمه "تخم مرغ" با اینکه انتهای کلمه "تخم" علامت کسره وجود ندارد ولی باید با کسره تلفظ شود. برای این زبان‌ها وجود فرهنگ لغت بسیار ضروری‌تر می‌باشد و از قوانین فقط برای تولید تلفظ کلمات غیر معمولی یا کلماتی که در فرهنگ لغت نیستند، استفاده می‌شود. تعداد این قوانین برای هر زبان معمولا زیاد می‌باشد و نوشتن قوانینی که همه موارد و استثناها را دربر داشته باشد وقت‌گیر و گاهی اوقات غیر ممکن است. معمولا به جای نوشتن قوانین از روش‌های داده‌گرا استفاده می‌کنند که در این روش‌ها، قوانین به طور مستقیم از روی داده‌ها، بدست می‌آید.

یکی از ابتدایی‌ترین پروژه‌هایی که در این زمینه انجام گرفته است، NetTalk می‌باشد، این پروژه برای تبدیل متن انگلیسی به تلفظش بکار می‌رود، بدین منظور از یک شبکه عصبی برای یادگیری قوانین تبدیل متن به واج استفاده می‌نماید [Rosenberg 1987]. این سیستم، یک شبکه عصبی سه لایه از نوع MLP[6] می‌باشد. کلمه به لایه ورودی داده می‌شود. ورودی یک پنجره لغزان هفت حرفی می‌باشد، کلمات روی پنجره قرار می‌گیرند و حرف به حرف حرکت می‌کنند تا کل حروف کلمه در مرکز پنجره (در موقعیت چهارم) قرار گیرد یعنی در هر لحظه هفت حرف به شبکه داده می‌شود. خروجی شبکه عبارتست از واج معادل با حرف وسط (حرف چهارم) از پنجره ورودی. بعد از اینکه تلفظ یک کلمه به طور کامل، تولید گردید وزن‌های شبکه بروزرسانی می‌شوند.

در[Fabio 2001] برای کاهش تعداد ناسازگاری‌های NetTalk راه حل خاصی ارایه گردیده است که در آن، از شبکه عصبی به صورت دو مرحله‌ای استفاده می‌شود، در مرحله اول شبکه عصبی بین حالت‌های یک واجی و دو واجی تمایز قائل می‌شود (حالت دو واجی حالتیست که، یک حرف به دو واج نگاشت می‌شود) در مرحله دوم، دو شبکه عصبی به صورت موازی عمل می‌کنند تا یک یا دو واج را به طور مجزا شناسایی کنند. به جای استفاده از پنجره متقارن موقعیت مرکزی[7] از پنجره با

[3] text-to-phoneme
[4] grapheme-to-phoneme
[5] rule-based
[6] MultiLayer Perceptron
[7] central position

موقعیت غیر متقارن(SPAW)[8]، استفاده می‌شود در این حالت تعداد حروف سمت چپ و سمت راست حرفی که باید به واج نگاشته شود یکسان نیست. تعیین اندازه مناسب برای پنجره موجب می‌شود تا کمترین ناسازگاری در خروجی ایجاد شود. آزمایش‌ها نشان داده که پنجره SPAW نسبت به پنجره موقعیت مرکز، تعداد ناسازگاری کمتری دارد. روش دو مرحله‌ای برای تبدیل متن فارسی به معادل واجی آن مناسب می‌باشد، زیرا اکثر اوقات یک حرف فارسی به دو واج نگاشت داده می‌شود. (در فارسی واکه‌ها به دو دسته کوتاه و بلند تقسیم می‌شوند و اغلب واکه‌های کوتاه در نگارش املایی نوشته نمی‌شوند) مثلا کلمه "سبز" به معادل واجی /sabz/ تبدیل می‌شود، البته برای زبان فارسی، در مرحله اول باید مشخص شود که واج به صفر، یک یا دو واج نگاشت می‌شود. در شکل (۱-۲) نحوه نگاشت از املا به واج برای کلمه "خواهر" آورده شده است.

خ x → (یک واج)
و - → (صفر واج)
ا A → (یک واج)
ه ha → (دو واج)
ر r → (یک واج)

شکل۱-۲- تبدیل املا به واج برای کلمه "خواهر"

به جای شبکه عصبی می‌توان از درخت طبقه‌بندی و رگرسیون [Aravind 1998] نیز استفاده نمود.

۱ ۱ ۵ -تلفظ اسامی خاص

بخش عظیمی از کلمات جدید، که یک سیستم تبدیل متن به گفتار با آن مواجه می‌شود از دسته اسامی خاص می‌باشد. تهیه یک فرهنگ لغت از کلیه اسامی خاص کار مشکلی است و معمولا قوانین حرف به صدا برای بسیاری از اسامی خوب عمل نمی-کنند و بسیاری از اسامی دارای قانون خاصی برای تبدیل حرف به صدا نیستند. لذا بهتر است تکنیک‌های خودکار برای تلفظ این اسامی را بهبود ببخشیم. از تکنیک‌های خودکار می‌توان به شبکه عصبی، درخت تصمیم گیری و... اشاره کرد. اخیرا از درخت‌های تصمیم‌گیری برای تولید اتوماتیک تلفظ اسامی به میزان زیادی استفاده شده است. برای ساخت درخت تصمیم-گیری مناسب که تلفظ درست اسامی را تولید کند ابتدا یک مجموعه بزرگ داده یادگیری فراهم می‌کنند. این مجموعه باید نمایشی از کل فضای مسئله باشد. برای تولید تلفظ، درخت تصمیم‌گیری با استفاده از مجموعه جفت‌های اسم-تلفظ آموزش داده می‌شود. بدین منظور از یک پنجره n حرفی استفاده می‌شود و کلمه مربوطه روی پنجره حرکت می‌کند، هر موقعیت کلمه در پنجره یک دنباله آموزشی برای درخت فراهم می‌کند و درخت باید یاد بگیرد که آن دنباله ورودی را به واج معادل با حرف وسط نگاشت دهد.

۱ ۲ -*تولید کننده نوای گفتار*

نوا، خصوصیت اصلی سیگنال گفتار می‌باشد که مرتبط با تغییرات گام، شدت و دیرش هجاها در صدا می‌باشد. تولید نوای گفتار مناسب یکی از مهمترین عوامل در تولید گفتار با کیفیت مناسب می‌باشد. نوا انتقال دهنده معنا و ساختار گفتار است. مشخص می‌کند که جمله خبری یا پرسشی می‌باشد و بین جملات پشت سر هم ارتباط بر قرار می‌کند. تخمین نوا مشکل است زیرا متن ورودی به سیستم TTS مستقیما هیچ اطلاعاتی در مورد نوا ارائه نمی‌دهد و باید با توجه به ساختار متن و ویژگی‌های زبانشانسی، نوا را برای آن تخمین زد. پارامترهای نوا عبارتند از دیرش، انرژی و تغییرات گام که برای تخمین مقادیر آنها، روش‌های مختلفی از قانون‌گرا گرفته تا داده‌گرا ارائه شده است. یکی از اولین سیستم‌های قانون‌گرا برای تولید نوا، سیستم

[8] Second Position Asymmetric Windowing-SPAW

کلات [Klatt1987] می‌باشد که در آن با استفاده از یکسری قوانین، پارامترهای نوا برای متن ورودی تولید می‌شود. به دلیل پیچیده بودن مسئله تعیین نوا و نامشخص بودن کلیه عوامل تاثیر گذار در نوا، امروزه روش‌های داده‌گرا جای روش‌های قانون-گرا را گرفته‌اند که در این روش‌ها، از روی داده‌های در دسترس، مدل‌های تولید نوا را ایجاد نموده و با استفاده از مدل‌های ایجاد شده پارامترهای نوا برای هر متن ورودی تخمین زده می‌شود.

هدف از مدلسازی کشش، پیدا کردن یک رابطه، بین عوامل موثر در کشش می‌باشد به‌طوری‌که مدل دارای دقت بالا و قابلیت تفسیر باشد. به دلیل اینکه تولید کلیه قوانین برای تخمین دیرش، بسیار سخت و پیچیده می باشند لذا امروزه از روش‌های داده‌گرا و آماری برای تخمین دیرش استفاده می شود. روش‌های آماری را می‌توان به دو دسته مدل‌های پارامتری و غیر پارامتری تقسیم‌بندی نمود. در مدل‌های پارامتری، ساختار پردازش پارامترهای ورودی از قبل مشخص است. نمونه این سیستم‌ها، مدل جمع-حاصلضرب‌ها[9]، مدل‌های خطی تعمیم یافته[10]، مدل‌های جمعی[11] و ضربی[12] می‌باشند. در مدل جمع-حاصلضرب‌ها، دیرش واحدها به صورت جمع یکسری عوامل نمایش داده می‌شود و تعامل بین هر ویژگی با دیگر ویژگی‌ها، به صورت یک جمله حاصلضرب مدل می‌شود[Santen 1994]. در مدل‌های غیر پارامتری، ساختار مدل به طور اتوماتیک با استفاده از نمونه‌های آموزشی بدست می‌آید. از روش‌های غیر پارامتری که در تولید دیرش بکار رفته‌اند می‌توان به شبکه‌عصبی [Campbell 1990] [Riedi 1995]،[Rumelhart 1986] ماشین پشتیبان بردار[Sreenivasa 2005]، درخت تصمیم‌گیری و تقریب [Sridhar 2004]،[Riley 1992]، منحنی‌های متعدد تقریب انطباقی[13] [Riedi 1997] و شبکه‌های باور بیزی [Goubanova 2000] اشاره نمود. در همه این روش‌ها، مسئله اصلی پیدا کردن عوامل تاثیرگذار روی دیرش می‌باشد. هر چه عوامل موثر را بیشتر بشناسیم دقت مدل‌های تولید کننده دیرش بالاتر می‌رود.

منحنی گام اصلی‌ترین بخش نوا می‌باشد که برای تخمین آن روش‌های مختلفی ارائه شده است. هیچ‌کدام از روش‌های مطرح شده قادر به تخمین ۱۰۰ درصد دقیق منحنی گام نیستند. از روش‌های پرکاربرد آن، می‌توان به مدل فوجی‌ساکی و تیلت اشاره نمود. مدل فوجی‌ساکی برای کنتور گام، دو بخش دستورات تکیه و عبارت را فرض می‌کند. دستورات تکیه تغییرات محلی کنتورگام و دستورات عبارت، تغییرات عمومی کنتور گام را مد نظر قرار می‌دهند. استخراج پارامترهای فوجی‌ساکی برای کنتور گام، یکی از مسائل حل نشده در مدل فوجی‌ساکی می‌باشد. تخمین موقعیت دستورات تکیه و عبارت و پارامترهای مربوط به آنها می‌تواند با استفاده از روش‌های یادگیری ماشین مانند شبکه عصبی، ماشین پشتیبان بردار و... انجام گیرد. مدل تیلت، رویدادهای آهنگی کنتور پیچ را دنبال می‌کند. برای هر رویداد یکسری پارامتر در نظر می‌گیرد که تخمین پارامترهای مدل تیلت برای یک متن داده شده با استفاده از روش‌های یادگیری ماشین می‌تواند انجام گیرد. در فصل سوم، روش‌های مختلفی که برای مدلسازی کنتور گام استفاده می‌شوند به طور کامل، شرح داده می‌شوند.

یکی دیگر از واحدهای اصلی نوای گفتار، انرژی گفتار می‌باشد که در ارسال مفهوم ضمنی متن، موثر می‌باشد. معمولا گوینده مواضع تکیه و محل‌های مورد تاکید در متن را با انرژی بیشتری قرائت می‌کند. انرژی دارای نام‌های دیگری ، مانند شدت، بلندی و دامنه نیز، می‌باشد. انرژی گفتار نیز به صورت یک منحنی در نظر گرفته می‌شود و برای تخمین آن یکسری روش ارائه شده است. در [Mixdorff 2003] نویسندگان مقاله، یک مدل قابل آموزش مجتمع را ارائه داده‌اند که درآن برای هر هجا، مقادیر انرژی، دیرش و پارامترهای مدل گام، تخمین زده می‌شود. در این سیستم، برای هر هجا یک مقدار انرژی تولید می‌شود و انرژی به صورت یک کنتور در نظر گرفته نمی‌شود. این مقدار انرژی میانگین کل انرژی فریم‌های آن هجا می‌باشد و برای تخمین مقدار آن، از شبکه عصبی استفاده می‌شود. در [Jokisch 2003] نویسندگان مقاله، تحقیقی در مورد عوامل موثر در تولید انرژی گفتار زبان آلمانی انجام داده‌اند. بدین منظور میزان همبستگی بین یکسری عوامل و مقدار انرژی یک واج را محاسبه نموده‌اند و بدین ترتیب، اهمیت ویژگی‌های مختلف در تولید انرژی را بدست آورده‌اند. در [Kyoung 2004] منحنی انرژی را به صورت یکسری تکه منحنی در محدوده واج در نظر گرفته و سعی نموده است ضرائب چندجمله‌ای مربوط به هر

---

[9] Sum-of-Products
[10] generalized linear
[11] additive
[12] multiplicative
[13] Multivariate Adaptive Regression Splines

منحنی را با استفاده از شبکه عصبی، تخمین بزند. در [Ghaemmaghami 2004]برای تولید لگاریتم منحنی پیچ و انرژی گفتار فارسی، از منحنی‌های تکه‌تکه در سطح هجا، استفاده شده و برای نمایش هر منحنی، از ضرائب لژاندر استفاده کرده است و سپس با استفاده از یک شبکه عصبی بازگشتی[14]، این ضرائب را برای هر هجا تخمین زده است.

از دیگر پارامترهای نوا می‌توان به مکث اشاره نمود، تخمین درست موقعیت مکث، روی معنای جمله تاثیر می‌گذارد و اگر مکان مکث را درست تعیین نکنیم معنی جمله اشتباه تفسیر می‌شود. پیش‌بینی مکث و مدت زمان آن یک مسئله چند متغیره می‌باشد که هنوز به طور کامل مدل نشده است. از عواملی که در مدت آن موثر هستند می‌توان به سرعت بیان گفتار (تعداد هجای بیان شده در واحد زمان)، حالت گوینده و سخن، پیچیدگی ساختار جمله، دیرش واج قبلی و... اشاره کرد. مثلا اگر طول عبارت‌های قبل و بعد از مکث کوتاه باشند احتمال اینکه مدت زمان مکث نیز کوتاه باشد زیاد است همچنین گوینده در هنگام ادای جمله‌های طولانی مکث می‌کند.

یکی از مشکلاتی که در زمینه مدل کردن نوا وجود دارد تاثیر متقابل پارامترهای نوا بر یکدیگر می‌باشد. معمولا بخش-های مختلف نوا یعنی دیرش، گام و انرژی را مستقل از هم در نظر می‌گیرند و سعی در تخمین هر کدام از این موارد به طور جداگانه دارند در صورتیکه این عوامل بر یکدیگر تاثیر متقابل می‌گذارند یعنی دیرش روی فرکانس پایه و فرکانس پایه روی دیرش تاثیر گذار است و دقت تخمین هر یک به دقت تخمین دیگری وابسته است که این خود یک دور باطل ایجاد می‌کند. بیشتر سیستم‌ها این دو را مستقل از هم در نظر می‌گیرند.

۳ ۱ -سنتز گفتار

به طور کلی انواع سنتز کننده‌های گفتار به چهار نوع سنتزکننده مفصلی، فرمنتی، پیوندی و ریاضیاتی تقسیم می‌شوند. روش های مفصلی سعی می‌کنند سیستم تولید گفتار انسان را به طور مستقیم مدل کنند. سنتز کننده‌های فرمنتی فرکانس تشدید سیگنال گفتار یا تابع تبدیل مجرای گفتار را بر اساس مدل فیلتر منبع، مدل می‌کنند و لذا هر دو روش، به اطلاعات زیادی در مورد نحوه تولید گفتار در انسان نیاز دارند. سنتز کننده‌های پیوندی، نمونه صوتی که قبلا ذخیره شده‌اند را برای سنتز بکار می‌گیرند و معمولا سنتز با استفاده از شکل موج یا استفاده از ضرائب پیشگویی خطی صورت می‌گیرد.

۱ ۳ ۱ -سنتز کننده مفصلی[15]

این روش سعی می‌کند فرایندهای فیزیکی که در هنگام تولید گفتار رخ می‌دهد را مدل کند. این روش‌ها به طور مستقیم یا غیر مستقیم موقعیت هر عضو گویایی را نشان می‌دهند. مدل مستقیم سعی می‌کند تا به طور صریح موقعیت مکانی هرعضو گویایی را ذخیره کند و روش غیر مستقیم فقط شکل کلی مجرای گفتار را در نظر می‌گیرد. بهترین و انعطاف‌پذیرترین مدل برای سنتز گفتار استفاده از این روش می‌باشد که سعی می‌کند فرایند فیزیکی تولید گفتار در انسان را مدل کند. اگر داده‌های مورد نیاز این مدل به اندازه کافی باشد صدای بسیار طبیعی را تولید می‌کند. پارامترهای مورد نیاز این مدل موقعیت فیزیکی هر عضو گویایی (در روش سنتز مستقیم) و یا شکل مجرای گفتار (در سنتز غیر مستقیم) می‌باشد. [Kroger 1992, Rahim 1993]

مشکل این روش کم بودن داده‌های تولیدی از گفتار طبیعی می‌باشد. در ابتدا از تصاویر اشعه x برای مشخص کردن تغییرات مجرای گفتار در طی تولید گفتار استفاده می‌شد این روش خیلی موفق نبوده است، زیرا از روی عکس دو بعدی باید ضرایب انعکاسی یک محیط سه بعدی را استخراج کرد[Klatt 1987]. امروزه به جای آن از تصاویرMRI استفاده می‌شود. همچنین می‌توان در نقاطی از سیستم تولید گفتار که می‌خواهیم نحوه حرکت آنها در حین تولید گفتار را بفهمیم یکسری سنسور قرار دهیم، در [Kello 2004] با قرار دادن سنسورهایی در محل‌های مختلف مانند گوشه لب بالایی، گوشه لب پایینی، دندان پیش بالا، دندان پیش پایین، نوک زبان، تیغه زبان، گوشه‌های زبان و سقف دهان با یک فرکانس نمونه‌برداری مشخص موقعیت این سنسورها در هنگام تولید گفتار توسط شخص، بدست آمده و از روی موقعیت آنها طی یکسری پردازش‌ها که انجام

[14] Recurrent neural network
[15] articulatory