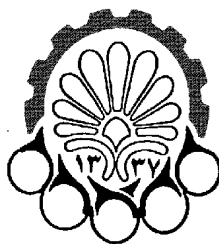


بسم الله الرحمن الرحيم



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

بسمه تعالی



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	<input type="checkbox"/>	بورسیه	<input type="checkbox"/>	معادل	<input type="checkbox"/>		
شماره دانشجویی:	۸۳۱۳۱۲۰۴	دانشکده:	مهندسی کامپیوتر	رشته تحصیلی:	هوش مصنوعی				
نام و نام خانوادگی استاد راهنما:	دکتر محمد مهدی همایونیور								
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:	<input type="checkbox"/> کارشناسی ارشد <input type="checkbox"/> دکترا	کاربردی	<input type="checkbox"/>	بنیادی	<input type="checkbox"/>	توسعه ای	<input type="checkbox"/>	نظری	<input type="checkbox"/>
تاریخ شروع:	بهمن ۱۳۸۴	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 پیشنهادات 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 همایون پور، این پایان نامه به اتمام رسیده است، بر خود واجب می دانم از زحمات ایشان، که با راهنمایی ها و کمک هایشان 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 ایران نژاد به خاطر ارائه نقطه نظرهای مفیدشان در زمینه بخش های 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 نمودم، عرضه می دارم.

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 به طوری که در اکثر موارد، کارایی آن در آموزش و تست، تقریباً یکسان می‌باشد. شبکه عصبی در محیط‌های نوپزی خیلی خوب عمل می‌کند اما امکان تفسیر خروجی ندارد.

نوای گفتار شامل دیرش، فرکانس پایه و انرژی آن می‌باشد که معمولاً 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 فوجی‌ساک، تیلت و منحنی‌های قطعه‌قطعه استفاده شد. روش فوجی‌ساک 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 هجاهای عبارتهای نوایی گفتار تخمین زده می‌شود و با استفاده از این پارامترها، منحنی گام با بکارگیری فرمول فوجی‌ساک، 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 فوجی‌ساک نمی‌باشد.

مدل تیلت، منحنی گام را به صورت دنباله‌ای از رویدادهای آهنگین فرض می‌کند. رویدادهای اصلی شامل تکیه زیرومی (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 کنترلر F_0 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

۲	۱- سیستم تبدیل متن به گفتار و اجزاء آن.....
۲	۱-۱- تحلیلگر متن.....
۳	۱-۱-۱- نرمال سازی متن.....
۳	۱-۱-۲- رفع ابهام از کلمات همشکل (هم نویسه ها).....
۳	۱-۱-۳- تحلیل تکواژشناسی.....
۴	۱-۱-۴- تبدیل حرف به صدا.....
۵	۱-۱-۵- تلفظ اسامی خاص.....
۵	۱-۲- تولید کننده نوای گفتار.....
۷	۱-۳- سنتز گفتار.....
۷	۱-۳-۱- سنتز کننده مفصلی.....
۸	۱-۳-۲- سنتز کننده فرمندی.....
۸	۱-۳-۳- سنتز کننده پیوندی.....
۹	۱-۳-۴- سنتز کننده های مبتنی بر تکنیک های ریاضی.....
۱۰	۱-۴- خلاصه.....
۱۲	۲- نوای گفتار.....
۱۲	۲-۱- تعریف و ماهیت فیزیکی تکیه.....
۱۲	۲-۲- درجات تکیه.....
۱۲	۲-۳- جای تکیه در واژه های زبان فارسی.....
۱۲	۲-۳-۱- تکیه اسم.....
۱۳	۲-۳-۲- تکیه صفت.....
۱۳	۲-۳-۳- تکیه عدد.....
۱۳	۲-۳-۴- تکیه ضمیر.....
۱۳	۲-۳-۵- تکیه فعل.....
۱۴	۲-۳-۶- تکیه قید.....
۱۴	۲-۳-۷- تکیه حرف اضافه.....
۱۴	۲-۳-۸- تکیه حرف ربط.....
۱۴	۲-۳-۹- تکیه اصوات.....
۱۵	۲-۴- وضع تکیه در جمله.....
۱۵	۲-۵- واحد آهنگین.....
۱۶	۲-۵-۱- ساختمان واحد آهنگین.....
۱۶	۲-۵-۲- طرح آهنگین جمله های خبری.....
۱۶	۲-۵-۳- طرح آهنگین جمله های پرسشی.....
۱۷	۲-۵-۴- طرح آهنگین جمله های امری.....
۱۷	۲-۵-۵- طرح آهنگین جمله های تعجیبی.....
۱۸	۲-۶- هجای هسته بر.....
۱۸	۲-۶-۱- محل هجای هسته بر.....
۱۸	۲-۷- ساخت گروه های نحوی.....

- ۱۸-۲-۷-۱ ساخت گروه اسمی.....
- ۱۹-۲-۷-۲ ساخت گروه حرف اضافه دار.....
- ۱۹-۲-۷-۳ ساخت گروه صفتی.....
- ۲۰-۲-۷-۴ ساخت گروه قیدی.....
- ۲۰-۲-۷-۵ ساخت گروه فعلی.....
- ۲۰-۲-۸-۸ مکث.....
- ۲۱-۲-۸-۱ آیا درنگ در زبان فارسی، عامل تمایز دهنده معنایی است؟.....
- ۲۱-۲-۹-۹ خلاصه.....
- ۲۳-۳ معرفی روش های مختلف مدلسازی آهنگ.....
- ۲۴-۳-۱ نظریه های زبانشناسی مرتبط با آهنگ.....
- ۲۴-۳-۲ مدل Fujisaki.....
- ۲۵-۳-۲-۱ دستور عبارت.....
- ۲۶-۳-۲-۲ دستور تکیه.....
- ۲۷-۳-۲-۳ استخراج اتوماتیک پارامترهای مدل فوجی ساکی از روی کنتور پیچ.....
- ۲۸-۳-۲-۴ مشکلات کنتور پیچ و پیش پردازش آن.....
- ۲۹-۳-۲-۵ فیلتر بالا گذر و جدا کردن اجزاء.....
- ۲۹-۳-۲-۶ مقداردهی پارامترهای مدل فوجی ساکی.....
- ۳۰-۳-۲-۷ تخمین پارامترهای مدل فوجی ساکی با استفاده از روش های ترکیبی داده گرا/قانون گرا.....
- ۳۲-۳-۲-۸ استخراج پارامترهای کنتور گام همزمان با تخمین آنها.....
- ۳۳-۳-۲-۹ مشکلات فوجی ساکی.....
- ۳۴-۳-۳-۳ مدل تیلت.....
- ۳۴-۳-۳-۱ معرفی تیلت.....
- ۳۶-۳-۳-۲ برچسب گذاری رویدادها.....
- ۳۶-۳-۴-۴ منحنی های قطعه قطعه در مدلسازی کنتور گام.....
- ۳۶-۳-۴-۱ استخراج ضرائب چندجمله ای برای قطعه منحنی ها.....
- ۳۷-۳-۵-۵ مدلسازی انرژی گفتار.....
- ۳۷-۳-۵-۱ منحنی های قطعه قطعه.....
- ۳۸-۳-۵-۲ استخراج انرژی گفتار.....
- ۳۸-۳-۶-۶ خلاصه.....
- ۴۱-۴-۴ روش های یادگیری ماشین.....
- ۴۱-۴-۱-۴ مدل مارس.....
- ۴۱-۴-۱-۱-۴ اساس روش مارس.....
- ۴۲-۴-۱-۲ مدلسازی پارامتری عمومی و غیر پارامتری محلی.....
- ۴۵-۴-۱-۳ توابع پایه مارس.....
- ۴۹-۴-۱-۴ شیوه تولید توابع پایه.....
- ۴۹-۴-۱-۵ متغیرهای از نوع سمبولیک (گروهی).....
- ۵۰-۴-۱-۶ ساخت مدل های تعاملی.....
- ۵۲-۴-۱-۷ مدل مارس.....
- ۵۲-۴-۱-۸ جلوگیری از برازش زیادی در مارس.....

- ۵۳..... ۹-۱-۴- پارامترهایی که باید تنظیم کنیم.....
- ۵۴..... ۱۰-۱-۴- مدت زمان اجرای الگوریتم مارس.....
- ۵۴..... ۱۱-۱-۴- تجزیه ANOVA.....
- ۵۵..... ۱۲-۱-۴- انتخاب کننده مدل مارس.....
- ۵۵..... ۱۳-۱-۴- الگوریتم مارس.....
- ۵۶..... ۲-۴- ماشین پشتیبان بردار.....
- ۵۶..... ۱-۲-۴- ماشین پشتیبان بردار طبقه بندی کننده.....
- ۵۶..... ۲-۲-۴- ماشین پشتیبان بردار تقریب زنده.....
- ۵۹..... ۳-۴- درخت طبقه بندی و رگرسیون.....
- ۵۹..... ۱-۳-۴- درخت طبقه بندی.....
- ۶۱..... ۲-۳-۴- روش های هرس کردن.....
- ۶۲..... ۳-۳-۴- نمونه های آموزشی با صفات فاقد مقدار.....
- ۶۲..... ۴-۳-۴- انواع یادگیری در درخت تصمیم گیری.....
- ۶۲..... ۵-۳-۴- تابع هدف.....
- ۶۲..... ۶-۳-۴- درخت رگرسیون.....
- ۶۳..... ۷-۳-۴- قابلیت تفسیر درخت CART.....
- ۶۳..... ۴-۴- شبکه عصبی.....
- ۶۴..... ۱-۴-۴- اندازه مجموعه آموزشی و تعداد وزن های شبکه عصبی.....
- ۶۴..... ۲-۴-۴- شبکه عصبی برای چه نوع مسئله هایی مناسب می باشد؟.....
- ۶۵..... ۳-۴-۴- جلوگیری از برازش زیادی در شبکه عصبی.....
- ۶۵..... ۵-۴-۴- الگوریتم ژنتیک.....
- ۶۵..... ۱-۵-۴- معرفی.....
- ۶۶..... ۲-۵-۴- کاربردهای الگوریتم ژنتیک.....
- ۶۷..... ۶-۴- خلاصه.....
- ۶۹..... ۵-۵- دادگان و ویژگی ها.....
- ۶۹..... ۱-۵-۱- دادگان مورد استفاده.....
- ۶۹..... ۲-۵-۲- ویژگی های بکار گرفته شده.....
- ۶۹..... ۳-۵-۳- عوامل تاثیرگذار بر دیرش واج.....
- ۷۰..... ۴-۵-۴- عوامل تاثیرگذار بر فرکانس پایه.....
- ۷۰..... ۵-۵-۵- مجموعه ویژگی ها.....
- ۷۱..... ۶-۵-۶- تحلیل متن.....
- ۷۳..... ۷-۵-۷- خلاصه.....
- ۷۵..... ۶-۶- پیاده سازی و نتایج آن.....
- ۷۵..... ۱-۶-۱- مدلسازی دیرش زمانی واج.....
- ۷۵..... ۱-۶-۱- توزیع مقدار دیرش نسبت به ویژگی ها.....
- ۷۶..... ۲-۶-۱- بکارگیری مارس در تخمین دیرش واج.....
- ۸۰..... ۳-۶-۱- بکارگیری شبکه عصبی در تخمین دیرش واج.....
- ۸۲..... ۴-۶-۱- بکارگیری ماشین پشتیبان بردار در تخمین دیرش واج.....
- ۸۴..... ۵-۶-۱- مقایسه نتایج مدلسازی های متفاوت.....

۸۴	۲-۶- مدلسازی منحنی گام.....
۸۵	۱-۲-۶- فوجی ساکی.....
۹۲	۲-۲-۶- تیلت.....
۹۸	۳-۲-۶- منحنی های قطعه قطعه.....
۱۰۷	۳-۶- مدلسازی انرژی گفتار.....
۱۰۷	۱-۳-۶- بررسی عوامل موثر در تولید انرژی.....
۱۱۲	۲-۳-۶- استخراج انرژی گفتار و تولید ضرائب چند جمله ای.....
۱۱۲	۳-۳-۶- شبکه عصبی در تخمین ضرائب چند جمله ای.....
۱۱۲	۴-۳-۶- ماشین پشتیبان بردار در تخمین ضرائب چند جمله ای.....
۱۱۲	۵-۳-۶- مارس در تخمین ضرائب چند جمله ای.....
۱۱۳	۶-۳-۶- مقایسه کارایی روش های مختلف در تولید کنتور انرژی.....
۱۱۵	۷-۳-۶- بررسی نتایج مارس.....
۱۱۶	۴-۶- ارزیابی سیستم پیاده سازی شده.....
۱۱۷	۱-۴-۶- ارزیابی کیفیت نوای تولید شده با استفاده از تست شنوایی.....
۱۱۹	۵-۶- خلاصه.....
۱۲۱	۷- نتیجه گیری.....
۱۲۲	۸- پیشنهاد.....
۱۲۵	۹- فهرست مراجع و منابع.....

فهرست شکل‌ها

- شکل ۱-۱- دیاگرام یک تحلیلگر متن ۲
- شکل ۲-۱- تبدیل املا به واج برای کلمه "خواهر" ۵
- شکل ۱-۳- دیاگرام مدل فوجی ساکی ۲۵
- شکل ۲-۳- پاسخ ضربه مکانیسم کنترل عبارت برای دستورات ورودی با اندازه Ap مختلف و آلفا برابر با ۲ ۲۶
- شکل ۳-۳- دستورات تکیه با مقادیر مختلف Aa ۲۷
- شکل ۴-۳- دستورات تکیه با مدت زمانهای مختلف ۲۷
- شکل ۵-۳- کنتور گام (بالا) و هموار شده درونیایی شده آن (پایین) ۲۹
- شکل ۶-۳- HFC (خطوط یکپارچه) و LFC (خطوط خط چین) که مقدار Fb از آن کم شده است. HFC ۲۹
- شکل ۷-۳- ناحیه بندی یک کنتور هموار شده به نواحی دارای گرادیان مثبت یا منفی ۳۰
- شکل ۸-۳- معماری ترکیب داده گرا و قانون گرا برای تولید پروژدی در یک سیستم TTS ۳۱
- شکل ۹-۳- تخمین پارامترهای فوجی ساکی با استفاده از شبکه عصبی ۳۲
- شکل ۱۰-۳- شمای سیستمهای تولید کنتور گام که در آن بخش استخراج پارامترها از بخش تخمین پارامترها جدا میباشد. ۳۲
- شکل ۱۱-۳- شمای مدل ترکیبی استخراج-تخمین که مراحل استخراج و تخمین پارامترها با هم در تعامل هستند. ۳۳
- شکل ۱۲-۳- منحنی گام، رویدادهای آهنگی آن و مرکز هجاها ۳۴
- شکل ۱۳-۳- مدل Tilt و پارامترهای آن ۳۴
- شکل ۱۴-۳- نمونه ای از کنتور پیچ یک واج و ضرائب چند جمله ای استخراج شده برای آن ۳۷
- شکل ۱۵-۳- نمونه ای از کنتور انرژی یک واج و ضرائب چندجمله ای استخراج شده برای آن ۳۸
- شکل ۱-۴- نمونه ای از تقریب خطی، که با ایجاد سه خط در فضای ورودی، تقریبی از خروجی بدست آمده است. ۴۲
- شکل ۲-۴- هموارسازی با استفاده از اپراتور میانه ۴۳
- شکل ۳-۴- منحنی تقریب بدست آمده (چپ) برای داده های شکل سمت راست، توسط مدل مارس ۴۳
- شکل ۴-۴- توزیع داده ها به منظور پیدا کردن knot ها ۴۴
- شکل ۵-۴- مراحل تقریب مارس برای داده های با توزیع شکل (۴-۴) ۴۵
- شکل ۶-۴- تعدادی از توابع پایه که برای داده x تعریف شده اند ۴۶
- شکل ۷-۴- توزیع مقدار MV ۴۷
- شکل ۸-۴- رابطه بین هر یک از متغیرهای ورودی با مقدار خروجی ۴۷
- شکل ۹-۴- تقریب MV از روی INDUS ، بالا: تقریب با استفاده از یک تابع پایه، پایین: تقریب با استفاده از دو تابع پایه ۴۸
- شکل ۱۰-۴- یک تابع پایه استاندارد (راست) و زوج آینه‌ای آن (چپ) ۴۹
- شکل ۱۱-۴- منحنی تقریب بدست آمده توسط الگوریتم مارس روی داده های مثال یک ۵۳
- شکل ۱۲-۴- چهار نوع تابع خطا که می‌توانند در ماشین پشتیبان بردار بکار گرفته شوند ۵۷
- شکل ۱۳-۴- پراکندگی داده ها برای طبقه بندی با استفاده از درخت طبقه بندی کننده ۶۰
- شکل ۱۴-۴- مراحل ساخت درخت دودویی برای طبقه بندی داده های شکل (۴-۱۳) ۶۰
- شکل ۱۵-۴- نمونه ای از درخت CART ساخته شده برای پیش بینی طول واحد ها ۶۳
- شکل ۱-۵- نحوه مشخص کردن گروه های نحوی برای جمله "محض خاطر خدا دست از وهم و خیال بردار" ۷۲
- شکل ۲-۵- نحوه مشخص کردن گروه های نحوی برای جمله "این نوع موتورها وزنشان کم است" ۷۳
- شکل ۱-۶- نحوه توزیع دیرش نسبت به بعضی از ویژگیهای تولید کننده دیرش ۷۶
- شکل ۲-۶- توزیع دیرش داده‌های گفتاری مورد استفاده ۷۶
- شکل ۳-۶- مثالی از آنالیز یک جمله فارسی که به صورت خبری (a) و پرسشی (b)، بیان شده است ۸۵
- شکل ۴-۶- استخراج اجزاء فوجی ساکی برای گفتار "boxl o bad xuyi ?az sefAte nApasandast." ۸۶

- شکل ۵-۶- استخراج اجزاء فوجی ساکی برای گفتار "boxl o bad xuyi ?az sefAte nApasandast"..... ۸۶
- شکل ۶-۶- یکی از حالات قرار گرفتن دستورات تکیه و هجاهای تکیه دار، در کنار هم..... ۸۷
- شکل ۷-۶- یکی از حالات قرار گرفتن دستورات تکیه و هجاهای تکیه دار، در کنار هم..... ۸۷
- شکل ۸-۶- یکی از حالات قرار گرفتن دستورات تکیه و هجاهای تکیه دار، در کنار هم..... ۸۷
- شکل ۹-۶- یکی از حالات قرار گرفتن دستورات عبارت، در کنار هم..... ۸۷
- شکل ۱۰-۶- یکی از حالات قرار گرفتن دستورات عبارت، در کنار هم..... ۸۷
- شکل ۱۱-۶- منحنی گام اصلی و سنتز شده آن به کمک فوجی ساکی برای جمله "قاضی گاوکش را عفو کرد"..... ۹۰
- شکل ۱۲-۶- منحنی گام اصلی و سنتز شده آن به کمک فوجی ساکی برای جمله "پیرمرد خیلی وقت است که رفته"..... ۹۱
- شکل ۱۳-۶- منحنی گام اصلی و سنتز شده آن به کمک فوجی ساکی برای "باید لنتها و شمعهای ماشین را عوض کنی"..... ۹۲
- شکل ۱۴-۶- منحنی گام اصلی و سنتز شده آن به کمک فوجی ساکی برای "محض خاطر خدا، دست از وهم و خیال بردار"..... ۹۲
- شکل ۱۵-۶- کنتور گام برای جمله پرسشی "مگر پیپ پر از توتون نیست؟" این کنتور در انتها، حالت کاهنده دارد..... ۹۶
- شکل ۱۶-۶- منحنی گام اصلی و سنتز شده آن به کمک تیلت برای جمله "قاضی گاوکش را عفو کرد"..... ۹۷
- شکل ۱۷-۶- منحنی گام اصلی و سنتز شده آن به کمک تیلت برای جمله "محض خاطر خدا، دست از، وهم و خیال بردار"..... ۹۸
- شکل ۱۸-۶- نمونه ای از یک منحنی پیچ که دارای فریم های صفر میباشد و نحوه اصلاح آن ۹۹
- شکل ۱۹-۶- نمونه هایی از منحنی هایی که دارای شکل پیچیده هستند و نحوه برازش یک منحنی درجه دوم روی آنها..... ۱۰۰
- شکل ۲۰-۶- نحوه پراکندگی اولین ضریب در معادله چند جمله ای درجه دوم..... ۱۰۰
- شکل ۲۱-۶- نحوه پراکندگی دومین ضریب در معادله چند جمله ای درجه دوم..... ۱۰۱
- شکل ۲۲-۶- نحوه پراکندگی سومین ضریب در معادله چند جمله ای درجه دوم..... ۱۰۱
- شکل ۲۳-۶- منحنی اصلی گام و سنتز شده آن به کمک ضرائب چندجمله ای، برای جمله "قاضی گاوکش را عفو کرد"..... ۱۰۵
- شکل ۲۴-۶- منحنی گام و سنتز شده آن به کمک ضرائب چندجمله ای، برای جمله "پیرمرد خیلی وقت است که رفته"..... ۱۰۶
- شکل ۲۵-۶- منحنی گام و سنتز شده آن به کمک ضرائب چندجمله ای، برای "باید لنتها و شمعهای ماشین را عوض کنی"..... ۱۰۶
- شکل ۲۶-۶- منحنی گام و سنتز شده آن به کمک ضرائب چندجمله ای، برای "محض خاطر خدا، دست از وهم و خیال بردار"..... ۱۰۷
- شکل ۲۷-۶- میزان همبستگی بین عوامل مختلف و میانگین انرژی یک واج..... ۱۰۸
- شکل ۲۸-۶- مقدار انرژی در بخش های مختلف هجا و در حالات مختلف فرکانس پایه ۱۰۹
- شکل ۲۹-۶- مقدار انرژی در حالات مختلف دیرش (کوتاه، متوسط، طولانی)..... ۱۰۹
- شکل ۳۰-۶- میانگین انرژی در واجهای مختلف و مقدار میانگین انرژی (مربوط به واج جاری)..... ۱۱۰
- شکل ۳۱-۶- نوع واج همسایه قبل و تاثیر آن روی انرژی..... ۱۱۱
- شکل ۳۲-۶- انرژی در قسمتهای مختلف هجا..... ۱۱۱
- شکل ۳۳-۶- منحنی اصلی انرژی و سنتز شده آن به کمک ضرائب چندجمله ای برای جمله "قاضی گاوکش را عفو کرد"..... ۱۱۳
- شکل ۳۴-۶- منحنی اصلی انرژی و سنتز شده آن به کمک ضرائب چندجمله ای برای "پیرمرد خیلی وقت است که رفته"..... ۱۱۴
- شکل ۳۵-۶- منحنی انرژی و سنتز شده آن به کمک ضرائب چندجمله ای برای "باید لنتها و شمعهای ماشین را عوض کنی"..... ۱۱۴
- شکل ۳۶-۶- منحنی انرژی و سنتز شده آن با ضرائب چندجمله ای برای "محض خاطر خدا دست از وهم و خیال بردار"..... ۱۱۵
- شکل ۳۷-۶- دیاگرام کلی سیستم پیاده سازی شده ۱۱۸

فهرست جدول‌ها

جدول ۱-۴- اطلاعات آماری مربوط به داده‌ها.....	۴۶
جدول ۲-۴- توابع پایه بدست آمده در مرحله نخست.....	۵۱
جدول ۳-۴- توابع پایه بدست آمده در مرحله دوم، تولید توابع پایه تعاملی.....	۵۱
جدول ۴-۴- نتیجه حاصل از تجزیه ANOVA.....	۵۴
جدول ۱-۵- ویژگیهای بکار گرفته شده برای بدست آوردن مدل دیرش.....	۷۰
جدول ۲-۵- ویژگیهای بکار گرفته شده به منظور تخمین دیرش همراه با واریانس هر یک از آنها.....	۷۲
جدول ۱-۶- مقادیر مختلف حداکثر درجه تعامل و تاثیر آن روی کارایی.....	۷۷
جدول ۲-۶- تجزیه anova مربوط به گروههای دارای یک یا دو متغیر.....	۷۸
جدول ۳-۶- درجه اهمیت متغیرها در تولید دیرش.....	۸۰
جدول ۴-۶- ویژگیهای بکار گرفته شده در آموزش شبکه عصبی و تعداد نورونهای اختصاصی به هر یک.....	۸۱
جدول ۵-۶- مدلسازی با استفاده از شبکه عصبی، با مجموعه ویژگیهای افزونه و مجموعه ویژگیهای کاهش یافته.....	۸۲
جدول ۶-۶- کارایی الگوریتم چندی سازی LBG با اندازه کتاب کدهای متفاوت در تخمین دیرش واج.....	۸۲
جدول ۷-۶- تعداد داده های موجود در هر خوشه و مقدار متوسط دیرش آنها.....	۸۳
جدول ۸-۶- دقت طبقه بندی کننده های پنجگانه.....	۸۳
جدول ۹-۶- مقایسه کارایی شبکه عصبی، مارس و ماشین پشتیبان بردار در مدلسازی دیرش واج گفتار فارسی.....	۸۴
جدول ۱۰-۶- مقایسه ضریب همبستگی روشهای مختلف مدلسازی، برای تخمین کنتور پیچ با استفاده از روش فوجی ساکی.....	۸۹
جدول ۱۱-۶- ضریب همبستگی بین مقدار واقعی پارامترهای تیلت و مقدار تخمین زده شده آنها.....	۹۳
جدول ۱۲-۶- میزان تاثیر عوامل مختلف در تخمین فرکانس پایه لحظه شروع رویداد.....	۹۳
جدول ۱۳-۶- میزان تاثیر عوامل مختلف در تخمین زمان شروع رویداد.....	۹۴
جدول ۱۴-۶- میزان تاثیر عوامل مختلف در تخمین اندازه رویداد.....	۹۴
جدول ۱۵-۶- میزان تاثیر عوامل مختلف در تخمین دیرش رویداد.....	۹۴
جدول ۱۶-۶- میزان تاثیر عوامل مختلف در تخمین عدد تیلت.....	۹۴
جدول ۱۷-۶- مقایسه ضریب همبستگی روشهای مختلف مدلسازی، برای تخمین کنتور پیچ با استفاده از روش تیلت.....	۹۵
جدول ۱۸-۶- مقایسه روشهای مختلف مدلسازی، در تخمین شکل انتهای کنتور پیچ.....	۹۶
جدول ۱۹-۶- ضریب همبستگی روشهای مختلف مدلسازی، برای تخمین کنتور پیچ با استفاده از روش چندجمله‌ای.....	۱۰۲
جدول ۲۰-۶- تجزیه anova برای سومین ضریب.....	۱۰۳
جدول ۲۱-۶- نتایج مدلسازیهای مختلف برای تخمین کنتور انرژی.....	۱۱۳
جدول ۲۲-۶- میزان تاثیر عوامل مختلف در تخمین اولین ضریب.....	۱۱۵
جدول ۲۳-۶- مقایسه روشهای مختلف مدلسازی، در تخمین شکل انتهای کنتور پیچ با استفاده از روش فوجی ساکی.....	۱۱۶
جدول ۲۴-۶- روشهای مختلف مدلسازی، در تخمین شکل انتهای کنتور پیچ با استفاده از روش منحنی های قطعه‌قطعه.....	۱۱۶
جدول ۲۵-۶- مقایسه روشهای مختلف مدلسازی، در تخمین شکل انتهای کنتور پیچ با استفاده از روش تیلت.....	۱۱۶
جدول ۲۶-۶- جملات بکار گرفته شده در تست MOS و امتیازهای دریافت شده برای هر یک.....	۱۱۸

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 ^۱TTS 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^۲ 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، هر کدام 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.



شکل ۱-۱- 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، کوتاه‌نوشته‌ها و... به متنی تبدیل می‌شود که در آن تمام عناصر غیر املائی مورد اشاره به معادل لغوی و بدون ابهامشان تبدیل شده باشند. تحلیل زبانشناسی 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

^۱ Text-To-Speech

^۲ 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 ۴ 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 در آن تمام عناصر غیر املائی 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 ۴ ۲ - رفع ابهام از کلمات هم‌شکل (هم نویسه ها)

یکی از مسائلی که در تحلیل 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 در جایگاه اسم باشد 'babr' و اگر در جایگاه 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 ۴ 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 "شهر" تو^۳ نیز به تلفظ آن اضافه می‌شود.

۱ ۴ - تبدیل حرف به صدا

مشخص کردن تلفظ یک کلمه، معمولا تبدیل متن به واج^۴ یا املا به واج^۴، نامیده می‌شود و فرایندی است که در آن 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^۵ 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^۶ 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^۷ از پنجره با

³ text-to-phoneme

⁴ grapheme-to-phoneme

⁵ rule-based

⁶ MultiLayer Perceptron

⁷ central position

موقعیت غیر متقارن (SPAW)^۸، 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 املائی نوشته نمی‌شوند) مثلاً 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)
ها	→	ha	(دو واج)
ر	→	r	(یک واج)

شکل ۱-۲- تبدیل املا به واج برای کلمه "خواهر"

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

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 معمولاً قوانین حرف به صدا برای بسیاری از اسامی خوب عمل نمی‌کنند 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 تصمیم‌گیری و... اشاره کرد. اخیراً از درخت‌های تصمیم‌گیری برای تولید اتوماتیک تلفظ اسامی به میزان زیادی استفاده شده است. برای ساخت درخت تصمیم‌گیری 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.

۱-۴- 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 برقرار می‌کند. تخمین نوا مشکل است زیرا متن ورودی به سیستم TTS مستقیماً هیچ اطلاعاتی در مورد نوا ارائه نمی‌دهد و باید با توجه به ساختار متن و ویژگی‌های زبانشناسی، 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 به طوری که مدل دارای دقت بالا و قابلیت تفسیر باشد. به دلیل اینکه تولید کلیه قوانین برای تخمین دیرش، بسیار سخت و پیچیده می‌باشند لذا امروزه از روش‌های داده‌گرا و آماری برای تخمین دیرش استفاده می‌شود. 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-حاصلضرب‌ها⁹، مدل‌های خطی تعمیم یافته¹⁰، مدل‌های جمعی¹¹ و ضربی¹² 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 انطباقی¹³ [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 کنترل گام، دو بخش دستورات تکیه و عبارت را فرض می‌کند. دستورات تکیه تغییرات محلی کنترل گام و دستورات عبارت، تغییرات عمومی کنترل گام را مد نظر قرار می‌دهند. استخراج پارامترهای فوجی‌ساکی برای کنترل 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 کنترل 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 کنترل 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. معمولاً 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 در آن 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 کنترل 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^{۱۴}، 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 بخش‌های 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.

۱-۳ سنتر گفتار

به طور کلی انواع سنتر کننده‌های گفتار به چهار نوع سنتر کننده مفصلی، فرمندی، پیوندی و ریاضیاتی تقسیم می‌شوند. روش‌های مفصلی سعی می‌کنند سیستم تولید گفتار انسان را به طور مستقیم مدل کنند. سنتر کننده‌های فرمندی 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. سنتر کننده‌های پیوندی، نمونه صوتی که قبلا ذخیره شده‌اند را برای سنتر بکار می‌گیرند و معمولا سنتر با استفاده از شکل موج یا استفاده از ضرائب پیشگویی خطی صورت می‌گیرد.

۱-۴ سنتر کننده مفصلی^{۱۵}

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 هر عضو گویایی را ذخیره کند و روش غیر مستقیم فقط شکل کلی مجرای گفتار را در نظر می‌گیرد. بهترین و انعطاف‌پذیرترین مدل برای سنتر 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 سنتر مستقیم) و یا شکل مجرای گفتار (در سنتر 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 تصاویر MRI 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