# مدل‌های آماری ترکیبی قابل آموزش با دادگان حجم متوسط برای تولید گونه‌های تلفظی کلمات در انواع ایستا و پویا

**بهرام وزیرنژاد** **فرشاد الماس‌گنج**

**دانشگاه صنعتی امیرکبیر، دانشکده مهندسی پزشکی، آزمایشگاه پردازش گفتار**

## چکیده

در گفتار پیوسته، کلمات به صورتهای گوناگونی تلفظ می‌شوند. وجود تنوعات تلفظی ایجاب می‌نماید واژگان سیستم بازشناسی گفتار پیوسته، حاوی گونه‌های تلفظی مناسب کلمات به همراه احتمال وقوع هریک از آنها باشد. با ایجاد چنین واژگانی می‌توان درصد دقت بازشناسی کلمات را افزایش داد. ابزار درخت تصمیم بطور وسیعی برای مدلسازی گونه‌های تلفظی کلمات و واحدهای زیر کلمه استفاده می‌شود. در صورتی که مدلسازی در سطح کلمه صورت گیرد، یعنی به ازای هر کلمه یک درخت تصمیم بطور خاص برای تعیین گونه‌های تلفظی آن تعلیم یابد، دادگان بسیار بزرگی شامل تمام کلمات موجود در واژگان با تعداد تکرار کافی مورد نیاز است. در این رویکرد علاوه بر نیاز به دادگان بسیار بزرگ، برای کلمات جدیدی که در دادگان آموزش موجود نباشند دچار مشکل می شویم بطوری که در عمل کاربرد آن برای سیستم های بازشناسی گفتار با واژگان خیلی بزرگ ناممکن می‌شود. در تحقیق حاضر برای حل این مسئله، درخت‌های تصمیم تعمیم یافته را طراحی نموده‌ایم. در این روش هر درخت مختص به یک کلمه نیست بلکه هر درخت مختص به گروهی از کلمات است که دارای ساختار واجی مشابه باشند. این درختها به صورتی کلی، توانایی پیش‌بینی آن نواحی از کلمه را که تبدیل، حذف و درج واج در آنها امکان داشته باشد، دارا هستند. بعد از یافتن نوع کلی تغییرات تلفظی کلمه از این طریق، قواعد تلفظی آماری، که از یک دادگان گفتاری با حجم متوسط استخراج شده‌اند و بطور کاملاً جزئی و دقیق مشخص می‌کنند هر واج در چه زمینهٔ واجی، چه تغییراتی خواهد کرد، به بخش‌هائی از کلمات که توسط درخت تصمیم مشخص شده‌اند، اعمال می‌شوند تا گونه‌های تلفظی کلمه تولید گردند. به این ترتیب یک مدل تلفظی شکل می‌گیرد که در عین عملکرد بر روی کل ساختار کلمه، نیازی به دادگان‌های بسیار بزرگ جهت آموزش ندارد. این مدل ترکیبی درخت تصمیم/ قاعده در حالت آموزش و استفاده به شکل ایستا، از ویژگی‌های ساختار واجی کلمه، محل تکیه در کلمه، احتمال وقوع کلمه در متون، و اطلاعات متنی واج‌ها استفاده می‌کند. در حالت آموزش و بکارگیری این مدل در شکل پویا، علاوه بر این ویژگی‌ها، نرخ گفتار نیز بطور همزمان به آن وارد می‌شود. با استفاده از گونه‌های تولید شده توسط این مدل، در حالت‌های ایستا و پویا، در واژگان سیستم بازشناسی گفتار پیوسته "شنوا"، به ترتیب ۸/۱٪ و ۱۰/۳٪ کاهش در نرخ خطای بازشناسی در سطح کلمه مشاهده گردید.

**کلمات کلیدی**: سیستم بازشناسی گفتار پیوسته، مدلسازی تنوعات تلفظی، گونه‌های تلفظی کلمات، واژگان.

## ۱- مقدمه:

اگر کلمات همیشه بطور یکسانی تلفظ می‌شدند؛ بازشناسی خودکار گفتار چندان مشکل نبود. ولی بدلایل گوناگون کلمات همواره به طرز یکسانی تلفظ نمی‌شوند[۱]. مهمترین عامل ایجاد تغییر در نحوه تلفظ کلمات در گفتار پیوسته، درهم فشردگی سیگنال آواها در این نوع گفتار است. در بازشناسی کلمات مجزا، گوینده بین کلمات مکث می‌نماید، که در نتیجه اثرات هم‌تولیدی[۱] آواها، بین کلمات کاهش می‌یابند. به علاوه در این نوع گفتار، گویندگان معمولا دقت بیشتری در ادای کلمات به‌خرج‌می‌دهند. اگرچه بازشناسی کلمات مجزا برای سیستم بازشناسی گفتار بسیار ساده‌تر از گفتار پیوسته است، ولی آدای کلمات بصورت مجزا برای انسان آسان نیست، و این نوع گفتار از حالت طبیعی دور است. لذا تلاش‌های زیادی صورت گرفته است تا سیستم‌های بازشناسی گفتار به فن‌آوری‌هایی مجهز گردند که بتوانند گفتار طبیعی را بازشناسی نمایند و این گونه است که هم‌اکنون تمرکز در تحقیقات بازشناسی گفتار در سطح دنیا بیشتر بر روی بازشناسی گفتار پیوسته و محاوره‌ای می‌باشد. واضح است که با تغییر شکل ورودی سیستم‌ها، از کلمات مجزا به گفتار پیوسته، میزان تنوعات تلفظی کلمات افزایش می‌یابد. ازآنجایی که وجود تنوعات تلفظی سبب بروز خطاهایی در بازشناسی گفتار خواهد شد، مدلسازی تنوعات تلفظی راهی برای افزایش کارآیی سیستم‌های بازشناسی گفتار می‌باشد. در گفتار پیوسته، برخلاف گفتار گسسته، انواع پدیده‌های هم‌تولیدی آواها می‌تواند در نواحی بین کلمات رخ دهد. در گفتار پیوسته با تغییر ویژگی‌های صوتی واج‌ها، میزان تبدیل، حذف و درج

واجها بیشتر رخ می‌دهند. میزان وقوع این پدیده‌ها به شیوهٔ صحبت گوینده وابسته است [۲]. شیوه صحبت گوینده، بر مبنای دوری یا نزدیکی به گفتار رسمی بیان می‌شود. هرچقدر گفتار گوینده غیر رسمی‌تر باشد، ساختار هجایی کلمات بیشتر تغییر میکند، نرخ گفتار بیشتر می‌شود، و تغییراتی در فرکانس پایه و قدرت سیگنال ایجاد می‌گردد [۳]. علاوه بر تنوعات تلفظی ناشی از شیوه بیان گوینده، عوامل مستقل دیگری نیز وجود دارند. مثلا گوینده می‌تواند از بین چند تلفظ رایج کلمه، یکی را انتخاب نماید بدون آنکه شیوه صحبت روی این فرآیند تاثیری داشته باشد. عامل مهم دیگر در ایجاد تنوعات تلفظی حالت احساسی گوینده است، مشخص گردیده است، گوینده‌ها تحت تاثیر مخاطب می‌باشند. در برخی از منابع حالت احساسی گوینده بعنوان عامل جداگانه‌ای برای بروز تنوعات تلقی نمی‌شود بلکه در همان دسته مربوط به شیوه صحبت درنظر گرفته می‌شود [۴]. تمام انواع تنوعات تلفظی که تاکنون ذکر گردیدند، بعنوان عوامل مستقل از گوینده در نظر گرفته می‌شوند. اگرچه میزان وجود این عوامل در بین گوینده‌های مختلف، متفاوت است. دسته دیگری از عوامل تاثیرگذار بر تنوعات تلفظی بین گوینده‌های یک زبان وجود دارد، که ناشی از گویش‌ها و لحن‌های مختلف در بیان آنها می‌باشد. گویش ولحن ویژه هر گوینده به عواملی مثل زادگاه، پیش‌زمینه اجتماعی-اقتصادی وی، سطح تحصیلات، جنسیت، سن و زبان مادری او بستگی دارد [۵]. علاوه بر عواملی که ذکر گردیدند، تنوعاتی در تلفظ که ناشی از وجود تفاوتهای آناتومیک در اندام‌های تولید صدا در افراد مختلف هستند، مشاهده می‌گردند. بعلاوه اخیرا تحقیقات نشان داده است، گوینده سعی می‌کند، کلماتی که مورد انتظار مخاطب نیست را به نحو واضح‌تر و شمرده‌تری بیان نماید، تا شنونده راحت‌تر بتواند این کلمات را بازشناسی نماید. لذا کلماتی که حائز این شرایط زبانی یا معنایی هستند، شامل تعداد کمتری از تبدیل، حذف و درج واج است[۶].

با توجه به نکات ذکرشده، کلمات تحت تاثیر عوامل گوناگون، به طرق متفاوتی چه از نظر مشخصات سیگنالی یا تنوعات واجی بیان می‌گردند، که این مساله دردسر بزرگی را برای سیستم بازشناسی گفتار ایجاد می‌نماید. مدلسازی تنوعات تلفظی برای اولین بار، در اوایل دهه ۱۹۷۰ مطرح گردید. در بسیاری از مقالات موجود در مجموعه مقالات کنفرانسهای بازشناسی گفتار IEEE در دهه ۷۰ این نکته مورد توجه قرار گرفته است. در همین مقالات لزوم استفاده از چندین تلفظ رایج برای هر کلمه، در واژگان[۲] سیستم بازشناسی که با استفاده از قواعد آواشناسی تولید می‌شوند، مورد توجه قرار گرفته است [۷]. در سالهای اخیر، تحقیقات وسیعی در این مقوله انجام گرفته است. اگرچه تنوعات تلفظی هم در ویژگی‌های صوتی آواها و هم در سطوح بالاتر مثل دنبالهٔ آوایی کلمات بروز می‌نماید، ولی بیشتر، تمرکز تحقیقات بر مدلسازی تنوعات در سطح ویژگی‌های آوایی سیگنالی گفتار بوده است. طی سالهای اخیر مدلسازی تنوعات تلفظی در سطح دنباله‌های واجی نیز تا حد زیادی مورد توجه قرار گرفته است. معمولا انتخاب عامل تاثیرگذار بر تنوعات تلفظی برای مدلسازی، بر اساس میزان تنوعاتی است که در گفتار تحت تاثیر این عامل ایجاد می‌شود و نیز براساس آثاری که این عامل بر عملکرد بازشناسی گفتار می‌گذارد، و در بسیاری از موارد، محققان عوامل تاثیرگذار در تنوعات تلفظی را از قبل انتخاب نمی‌کنند، بلکه آنها بر اساس تجزیه و تحلیل دادگان‌های گفتاری (در روش‌های دادگان محور) و بطور خودکار وارد مدل‌های تولید تنوعات گفتاری می‌شوند. این مطلب مقایسه نتایج این تحقیقات را که با راهکارهای گوناگون انجام‌شده‌اند، مشکل می‌نماید، و تفسیر نتایج را پیچیده میکند.

مروری بر مقالات انتشاریافته در ارتباط با مدلسازی تنوعات تلفظی درسیستم بازشناسی خودکار گفتار و عوامل ایجاد کننده آنها، نشان می‌دهد که معمولا محققین ارزیابی روش‌های خود را بر اساس میزان بهبود عملکرد سیستم بازشناسی گزارش می‌دهند. در بیشتر روشها محققان اقدام به وارد نمودن گونه‌های تلفظی[۳] متعدد برای هر کلمه در واژگان سیستم می‌نمایند. این روشها مبتنی بر این فرض است که بسیاری از تنوعات تلفظی در سطح ویژگی‌های سیگنالی آواهای بیان شده، منجر به تغییر کلی در واج‌های بازشناسی‌شده بصورت بروز پدیده‌هایی مثل تبدیل، حذف و درج واج، خواهند شد. استفاده از مدل‌های تلفظی در تولید واژگان سیستم باعث افزایش درصد دقت بازشناسی کلمات در سیستم می‌گردد. این بهبود ناشی از تطبیق صحیح‌تر واج‌های بازشناسی‌شده با ترکیب‌های واجی کلمات موجود در واژگان سیستم می‌باشد. بااینحال ارائه بیش از حد گونه‌های تلفظی در واژگان باعث افزایش سردرگمی[۴] در سیستم خواهد شد و این مساله می‌تواند صحت بازشناسی را کاهش دهد. بنابراین باید محدودیتهایی در پذیرش گونه‌های تلفظی قائل گردد و تنها از گونه‌هایی که احتمال وقوع بالا دارند در تولید واژگان سیستم بازشناسی استفاده شود [۸]. استفاده از واژگانی که حاوی گونه‌های تلفظی کلمات باشد، راه مناسبی برای بهبود عملکرد سیستم‌های بازشناسی گفتار است. تولید گونه‌های تلفظی کلمات بصورت دستی و مبتنی بر دانش زبانشناسی کار بسیار پیچیده‌ای است، و احتیاج به زمان زیادی دارد. در سالهای دهه ۹۰ تحقیقات زیادی در راستای تولید خودکار واژگان حاوی گونه‌های تلفظی کلمات صورت گرفته است. یکی از اولین تلاشها برای این منظور با استفاده از قواعد واجی بود، که در سیستم TBM انجام پذیرفت [۹]. در سالهای آغازین دههٔ ۱۹۹۰، با استفاده از پایگاه دادگانی نظیر TIMIT که در آن سیگنال گفتار واج‌نویسی شده است، محققانی نظیر راندولف[۵] ۱۹۹۰، ریلی[۶] ۱۹۹۱، ووترز[۷] و استولک[۸] ۱۹۹۴ تلاشهای زیادی برای مدلسازی تنوعات واجی نمودند[۱۰]. این تلاشها منجر به امکان تولید خودکار قواعد تلفظی شد. اما برای اینکار هنوز به واج‌نویسی فونتیک گفتار به‌صورت دستی احتیاج بود. در تحقیقات بعدی با استفاده از یک سیستم بازشناس واج، امکان استفاده از دادگان بزرگ بدون برچسب‌دهی واجی مهیا گردید. چنین تلاشهایی توسط اشمید[۹] در سال ۱۹۹۳،اسلوبودا[۱۰] ۱۹۹۵،امایی[۱۱] ۱۹۹۵ و هامفریز[۱۲] ۱۹۹۷ انجام گرفت [۱۰]. در سالهای اخیر، با تمرکز برای ایجاد سیستمهای بازشناسی گفتار محاوره‌ای با واژگان بزرگ، مدلسازی تلفظ بسیار اهمیت یافته است. چرا که تلفظ کلمات در گفتارمحاوره‌ای نسبت به گفتار روخوانی دارای تنوعات بسیار بیشتری است [۱۱].

در این قسمت به بررسی روشهای مختلف مدلسازی تنوعات تلفظی خواهیم پرداخت. مدلهای تلفظی می‌توانند قابلیت مدلسازی تنوعات تلفظی بین‌کلمه‌ای را داشته باشند و یا منحصر به مدلسازی تنوعات تلفظی داخل‌کلمه‌ای گردند. این مدل‌ها هم در سطح کلمه و هم واحدهای زیرکلمه طراحی شده‌اند. آموزش مدل‌های تلفظی یا مبتنی بر دانش زبانشناسی انجام می‌گیرد، و یا بصورت خودکار و با استفاده از دادگان آموزش مناسب صورت می‌پذیرد. پاسخ سؤالاتی نظیر اینکه: چه نوع تنوعات تلفظی را می‌خواهیم مدل نماییم؟ داده‌های مورد نیاز چگونه مهیا می‌شوند؟ آیا اطلاعات استخراج شده قالب یک مدل ارائه می‌گردند، یا بصورت ذکر گونه‌ها در واژگان ارائه می‌شوند؟ و اینکه تنوعات مدل شده در کدام قسمت سیستم بازشناسی گفتار مورد استفاده قرار خواهند گرفت، ویژگیهای مدل‌های تلفظی مورد طراحی را مشخص می‌نمایند. مدل‌های تلفظی به روشهای مختلفی از جمله جمع‌آوری قواعد تلفظی، آموزش درختهای تصمیم، استفاده از مدلهای مخفی مارکوف و شبکه‌های عصبی مصنوعی قابل ارائه می‌باشند [۱۲].

در مدلسازی با استفاده از قواعد تلفظی، هدف تولید گونه‌های تلفظی کلمات از طریق اعمال قواعد تلفظی روی دنباله‌های واجی مرجع کلمات می‌باشد. این روش مبتنی براین فرض است که تقریبا تمامی گونه‌های تلفظی کلمات از طریق اعمال قواعد تلفظی به دنباله واجی مرجع کلمه، قابل دستیابی می‌باشند. تعدادی زیادی از محققان از قواعد تلفظی در سیستم‌های بازشناسی خود استفاده نموده‌اند تا مشکل تنوعات تلفظی را حل نمایند. و از این طریق بهبودهای چشمگیری در خروجی سیستم‌ها گزارش نموده‌اند [۱۳، ۱۴، ۱۵و ۱۶]. قواعد تلفظی معمولا در قالب یک توصیف کیفی، تغییرات مجاز واجی را تعریف می‌نمایند و فاقد مقادیر کمّی احتمالات وقوع هستند. تاجمن[۱۳] این مساله را با تخمین احتمالات قواعد تلفظی از دادگان آموزش حل نمود [۱۷]. در برخی از روش‌های مدلسازی، گونه‌های تلفظی مستقیما از دادگان استخراج می‌گردند، اما روش‌های مبتنی بر قاعده، دارای مزایایی نسبت به این روش‌ها می‌باشند. اگر این قواعد

به اندازه کافی غنی و معتبر باشند، حتی می‌توان به تولید گونه‌های تلفظی کلماتی که در دادگان آموزش دیده نشده‌اند، پرداخت. این قابلیت در روش استخراج گونه‌ها بصورت مستقیم وجود ندارد. کرملی و مارتنز[۱۴] نیز برهمین اساس از قواعد استخراج شده بصورت خودکار برای تولید گونه‌های تمام کلمات استفاده نمودند [۱۳]. قواعد تلفظی یا بر اساس اطلاعات زبان‌شناسی ساخته می‌شوند و یا با استفاده از دادگان آموزش بصورت خودکار، استخراج گردند. در روش دوم بعد از همردیف‌سازی[۱۵] دنباله واجی مرجع و دنباله واجی بازشناسی شده کلمات و مقایسه دنباله‌های همردیف شده، قواعد به صورت خودکار استخراج می‌شوند. قواعد استخراج شده به روش دوم می‌توانند علاوه بر تنوعات آوایی دیده شده در سیگنال گفتار پیوسته، تغییرات واجی ناشی از ضعفهای سیستم بازشناسی آوا را نیز مدل کنند. این قواعد با توجه به دادگان آموزش شکل می‌گیرند و تعمیم آنها به دادگان آزمون، مستلزم هرس آماری آنها است. اینکار با توجه به معیارهای آماری ویا اعمال محدودیتهای مبتنی بر قواعد آواشناسی، قابل انجام است [۱۸]. استخراج قواعد تلفظی از دادگان با حجم متوسط متداول است. ضعف بارز قواعد تلفظی این است که آنها برای اینکه تشخیص دهند هر واج قابل تبدیل یا حذف هست یا خیر، تنها به چند واج اطراف آن در نسخه مرجع واجی توجه می‌نمایند و به ویژگی‌های دیگر نظیر ساختار کلی کلمه، موضع تکیه، نرخ گفتار و ... توجه نمی‌کنند.

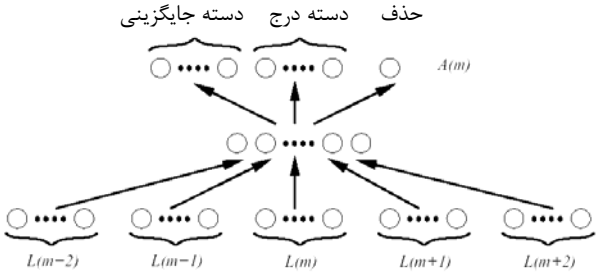
درخت تصمیم نیز می‌تواند ابزاری برای تولید گونه‌های تلفظی کلمات باشد. این‌کار از طریق بررسی ویژگی‌های دنباله واجی مرجع کلمه انجام می‌شود[۱۹]. درخت‌های تصمیم، ابزاری جهت طبقه‌بندی دادگان می‌باشند. در مرحله آموزش، درخت تصمیم ساختار خود را با استفاده از دادگان آموزش شکل می‌دهد [۲۰]. در ساخت یک درخت تصمیم چند نکته حائز اهمیت است: اول انتخاب سؤالاتی که به تفکیک اجزای دادگان آموزش می‌پردازند. برخی از الگوریتم‌ها می‌توانند بصورت خودکار سؤالات را طراحی نمایند. انتخاب سؤالات به‌گونه‌ای صورت می‌گیرد که حداکثر کاهش آنتروپی در هر مرحله تفکیک دادگان محقق شود. توسعه درخت تصمیم تا جایی انجام می‌گیرد که شرط توقف تقسیم گره‌ها تحقق یابد. این شرط وجود تعداد کمتر از حد آستانه عناصر موجود در آن گره و یا کاهش آنتروپی کلی درخت به زیر حد آستانه موردنظر می‌باشد. گره‌هایی که دیگر قابل تفکیک نیستند به گره‌های نهایی موسومند. بعد از آموزش وساخت درخت تصمیم، هرس آن به‌منظور تعمیم دهی آن به دادگان غیر از آموزش صورت می‌گیرد، چراکه ممکن است درخت تصمیم بیش از حد به دادگان آموزش تطبیق یابد. فوسلر[۱۶] با استفاده از درختهای تصمیم به مدلسازی تنوعات تلفظی موضعی مبادرت ورزید [۲۱]. همچنین واژگان پویای تلفظی را برای سیستم بازشناسی گفتار با استفاده از درخت تصمیم ارائه نمود [۲۲].

مدلهای مخفی مارکوف از دیگر ابزار مدلسازی تنوعات تلفظی است، بازشناسی گفتار در واقع، انتخاب بهترین کلمه با توجه به بردار ویژگی استخراج شده از سیگنال صوتی می‌باشد. کلمه بازشناسی شده باید در شرط زیر صدق نماید.

$$W^* = \arg\max_{[W \in L]} p(W \mid O) = \arg\max_{[W \in L]} p(O \mid W)\, p(W) \qquad (۱)$$

$W^*$ بهترین کلمه انتخاب‌شده است. $p(O \mid W)$ احتمال آن است که کلمه $W$ عضو واژگان $L$، به صورت دنباله واجی $O$ تلفظ گردد. $p(W)$ احتمال وقوع کلمه در حالت کلی است. محاسبه احتمال بیزین $p(O \mid W)p(W)$ با استفاده از مدلهای مخفی مارکوف صورت می‌گیرد. مدلهای مخفی مارکوف شامل یکسری گره بعنوان واج‌ها و یکسری احتمالات گذر بین گره‌ها یا حالات می‌باشند. با توجه به مدلهای مخفی مارکوف احتمالات $p(O \mid W)p(W)$ برای هر کلمه موجود در واژگان محاسبه می‌شود. محاسبه $p(O \mid W)$ برای هرکلمه با استفاده از الگوریتم‌های برنامه‌ریزی پویا و یا الگوریتم‌های ویتربی صورت می‌گیرد. وترز و استولک[۱۷] با استفاده از مدلهای مخفی مارکوف به مدلسازی واژگان چند تلفظی برای سیستم بازشناسی گفتار مستقل از گوینده پرداختند[۱۰]. استفاده از مدلهای مخفی مارکوف در سیستم‌هایی که طبقه شناسایی واج سیستم با واژگان سیستم ادغام شده باشند، مقدور است.

شبکه‌های عصبی روش دیگری برای مدلسازی تنوعات تلفظی می‌باشند که در برخی از منابع کارآیی آنها بهتر از روشهای آماری دیگر ذکر شده است. شبکه‌های عصبی قادر به پیش‌بینی دنباله واجی بازشناسی‌شده کلمات $\hat{P} = [\hat{p}_1, ..., \hat{p}_N]$ با در اختیار داشتن دنباله واجی مرجع کلمه، $P = [p_1, ..., p_N]$ می‌باشند. ویژگی‌های مورد استفاده شبکه‌عصبی شامل متن واجی یا دنباله واجی مرجع کلمه، به همراه دیگر ویژگی‌ها می‌باشند. فوکودا و ساجیساکی[۱۸] از گونه‌های تلفظی کلمات برای آموزش شبکه عصبی استفاده نمودند تا بتوانند گونه‌های تلفظی کلمات را پیش‌بینی نمایند [۲۳]. شکل ۱ ساختار شبکه مورد استفاده ایشان را نشان می دهد. تعداد نرونها در هر یک از دسته ها؛ (شامل دسته های نرونهای متنی در ورودی و دسته نرون های درج و دسته نرونهای جایگزینی) به تعداد واج های موجود در زبان است. در ورودی نرون مربوط به هر واج که واقع شده باشد ۱ و دیگر نرونهای دسته صفر فرض می شوند. و در خروجی نرون بیشینه در هر دسته برنده خواهد بود.




شکل ۱- ساختار شبکه عصبی مورد استفاده برای یادگیری تنوعات تلفظی [۶۳].

در کارهای جدیدتر از ویژگی‌های تمایزگر دیگری نظیر فرکانس پایه، جای تکیه کلمه و آهنگ بیان کلمه، بعنوان ویژگی‌های اضافی در ورودی شبکه عصبی استفاده می‌کنند، این ویژگی‌ها باعث پیش‌بینی بهتر نحوهٔ تبدیل واجها می‌شوند. چن و هازگاوا[۱۹] با تعریف چنین ویژگی‌هایی در ورودی شبکه عصبی، کارآیی سیستم خود را افزایش دادند [۲۴]. استفاده از شبکه‌های عصبی بعنوان مدلهای تلفظی مستلزم داشتن دادگان آموزش با حجم بسیار زیاد برای همگرا نمودن شبکه عصبی و طراحی بهینه برای ساختار شبکه می‌باشد. لذا با حجم دادگان متوسط استفاده از این روش رایج نیست.

علاوه بر روشهایی که ذکر گردیدند، روشهای غیر کلاسیک دیگری نظیر آنچه در مرجع [۲۵] توسط هازن و همکارانش گزارش شده است، دیده می شوند. هازن با استفاده از ماشین های حالت محدود به مدلسازی شرایط و نحوه ایجاد تنوعات تلفظی پرداخت. باید توجه داشت که در اکثر این روشها تشخیص نوع و نحوه انحراف تلفظ واقعی از گونه تلفظی مرجع بعنوان چارچوبی برای تعیین گونه های تلفظی میباشد. تعیین گونه های تلفظی سازگار و کارآ در بسیاری موارد دشوار است و در قالب قواعد تعریف شده نمیگنجد. در عین حال استخراج مجموعه قواعد عام تبدیل نوشتار به دنباله واج ها از واژگانی که حاوی کلمات و گونه های تلفظی کلمات باشد، دشوار است. داول و برنارد در [۲۶] این مسائل را با ایجاد آنچه "شبه واج[۲۰]" و "قواعد عام بازدارنده[۲۱]" معرفی کردند، مورد هدف قرار داده اند و در صدد حل آنها بر آمده اند.

در تعدادی از مراجعی که مربوط به کار روی مدلسازی و تولید گونه‌های تلفظی هستند، به اثر برخی از پارامترهای مهم و مؤثر در گونه‌های تلفظی پرداخته شده‌است. در اینجا به عنوان مثال به دو مورد مهم از آنها اشاره می‌نمائیم.

دیده می‌شود که کارآیی سیستم‌های بازشناسی گفتار پیوسته با افزایش نرخ گفتار ورودی کاهش می‌یابد، بطوریکه میزان خطا برای گفتار سریع ۲ تا ۳ برابر بیشتر از گفتار معمولی است [۲۷]. از دلایل اصلی افزایش میزان خطا برای گفتار سریع می‌توان به افزایش اثرات هم‌تولیدی آواها، اشاره نمود. مشخصات طیفی آواها در گفتار سریع با گفتار معمولی متفاوت هستند [۲۷]. علاوه بر این، گونه‌های تلفظی کلمات در گفتار سریع با گفتار عادی یکسان نیستند. دنباله‌های واجی کلمات در گفتار سریع نسبت به

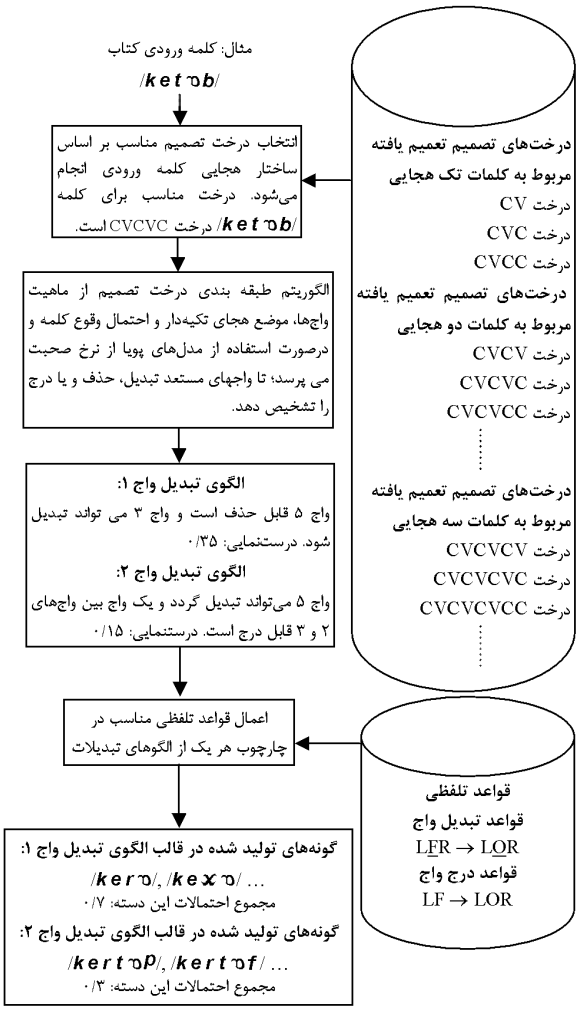
گفتار معمولی کوتاه‌تر هستند و برخی از واج‌ها تلفظ نمی‌شوند. در گفتار سریع طول زمانی واج‌ها کمتر از گفتار معمولی است. به منظور پایدارسازی سیستم‌های بازشناسی گفتار مجهز به فن‌آوری HMM، در برخورد با تنوعات در نرخ گفتار، احتمالات گذر بین واجها در مدلهای کلمه را برای گفتار سریع، اصلاح نموده‌اند. به این ترتیب مدلهای HMM کلمه برای گفتار سریع تطابق بیشتری با طول زمانی واجها در گفتار ورودی پیدا می‌کنند. ساختار مدلهای HMM کلمه نیز برای گفتار سریع که دارای تعداد واجهای کمتری در گونه‌های تلفظی کلمات می‌باشند اصلاح شده‌اند. در این مورد از گونه‌های تلفظی با طول دنباله واجی کوتاهتر برای طراحی مدلهای مربوط به گفتار سریع استفاده شده است. در مرجع [۲۷] از مدلهای HMM کلمه مناسب برای گفتار سریع و آهسته بصورت موازی استفاده شده است. همچنین تحقیقات اخیرا نشان داده است، میزان احتمال پیشگویی کلمه بر اساس مدل‌های زبانی با تنوعات تلفظی واجی مرتبط است [۶، ۲۲]. واضح‌ترین معیار برای میزان قابلیت پیش‌بینی کلمات، احتمال غیرشرطی وقوع کلمه یا $P(word)$، یا یک-تایی "unigram" است. در یک تحقیق ۵۶۱۸ گفته از ۱۰ کلمه پر تکرار انگلیسی شامل در نظر گرفته شده اند. سپس نشان داده شده است، که این کلمات دارای طول زمانی کمتر و حذف واکه‌ها می‌باشند. ضمنا t, d آخر در این کلمات، با احتمال زیادی قابل حذف هستند [۶].

## ۲- اصول روش پیاده سازی شده

مدل ترکیبی[۲۲] آماری درخت تصمیم/ قاعده که در این گزارش برای تولید گونه‌های تلفظی کلمات ارائه نموده‌ایم، به دو صورت ایستا[۲۳] و پویا[۲۴] قابل پیاده سازی است. مدل ایستا برای تولید گونه‌های تلفظی به ساختار واجی کلمه، موضع تکیه، احتمال وقوع و اطلاعات متنی واج‌ها توجه می‌نماید. مدل پویا علاوه بر این اطلاعات به نرخ گفتار نیز بعنوان یک ویژگی که در طول گفتار می‌تواند تغییر نماید، توجه می‌نماید. که در این مقاله کارآیی هر یک از آنها در بهبود عملکرد سیستم بازشناسی گفتار پیوسته با هم مقایسه خواهد شد. مدل ترکیبی آماری صرفنظر از اینکه ایستا یا پویا باشد، از ترکیب دو ابزار درخت تصمیم و قواعد تلفظی طراحی شده‌است.

قبل از این که به توضیح مدل ترکیبی ارائه شده بپردازیم، اشاره می شود که در تعدادی از تحقیقات انجام شده، برای مثال در کارهای [۲۲ و ۲۳]، برای مدلسازی تغییرات تلفظی برای هر کلمه بطور خاص یک درخت تصمیم طراحی شده است. در این رویکرد برای مثال، درخت تصمیم خاصی برای کلمه "کتاب" آموزش داده شده است. این درخت با توجه به ویژگی‌هایی خاص آن کلمه، نظیر موضع تکیه و احتمال وقوع، و همچنین با توجه به ساختار آوائی کلمات اطراف، گونه‌های تلفظی مناسبی را برای کلمه "کتاب" پیش‌بینی می‌نماید. در گزارش حاضر، بجای طراحی و آموزش یک درخت برای هر کلمه، که مشکلات فراوانی را از جهت فراهم آوردن دادهٔ آموزش کافی برای تمام کلمات فراهم می‌کند، یک درخت برای هر دسته کلمات با ساختار مشابه آوائی طراحی شده است. برای مثال از تمام کلماتی که ساختار و نحوه آرایش همخوانها (که با "C" نشان داده می‌شود) و واکه‌ها (که با "V" نشان داده می‌شود) در آنها بصورت "CVCVC" است، در آموزش یک درخت تصمیم واحد بهره می‌گیریم. در این رویکرد، اطلاعات تنوعات تلفظی ظاهر شده برای تمام کلمات آن دسته، در آموزش یک درخت واحد شرکت داده می‌شوند و مشکلی به صورت کمبود دادگان آموزش رخ نخواهد داد و علاوه بر عدم نیاز دادگان بسیار بزرگی که شامل تمام کلمات موجود در واژگان (به تعداد کافی)، برای کلمات جدید، نیاز به اضافه کردن دادگان جدید و فراهم آوردن دادگان اضافی نخواهد بود. به این ترتیب با استفاده از رویکرد پیشنهاد شده و به اشتراک گذاشتن دادگان مربوط به دسته ای از کلمات در تولید مدل مربوط به آن دسته، حجم دادگان مورد نیاز برای آموزش مدل‌ها به شدت کاهش می‌یابد. با توجه به این نکته که در این رویکرد، درخت‌های تصمیم بجای تعلق به یک کلمه، مربوط به گروه‌های کلمات با ساختار واجی مشابه هستند، به آنها لفظ "درخت‌های تصمیم تعمیم‌یافته" را نسبت می‌دهیم.

در این قسمت به تشریح مدل ترکیبی پیشنهادی برای تولید گونه‌های تلفظی کلمات می‌پردازیم. مراحل تولید گونه‌های تلفظی کلمات در شکل ۲ آمده اند. مدل ترکیبی آماری طی دو مرحله اصلی به تولید گونه‌های تلفظی کلمه ورودی می‌پردازد. در مرحله اول، درخت تصمیم موسوم به درخت تصمیم تعمیم‌یافته که متناظر با ساختار واجی کلمه ورودی عمل می‌کند، به پیش‌بینی اینکه کدام واج‌ها در کلمه قابل تبدیل یا حذف هستند و یا اینکه در کدام قسمت کلمه درج واج قابل انجام است، می‌پردازد. انتخاب درخت متناظر با کلمه ورودی براساس ساختار واجی یا نوع قرارگرفتن همخوان‌ها (که با C نشان داده شده‌اند) و واکه‌ها ( که با V نشان داده شده‌اند) در کلمه است. برای روشن‌تر شدن مسئله، مثالی در شکل ۲ آورده شده است. کلمه ورودی، "کتاب" با دنباله واجی مرجع /**k e t a b**/ است. این کلمه دو هجایی است و الگوی قرارگیری همخوان‌ها و واکه‌ها و در آن بصورت "CVCVC" می‌باشد. بنابراین درخت تصمیم تعمیم یافته متناظر با آن، درخت تصمیم تعلیم یافته برای ساختار واجی "CVCVC" است. این درخت، برای کلمات با ساختار واجی "CVCVC" آموزش دیده تا بتواند نحوه تبدیل، حذف و درج واج‌ها را پیش‌بینی نماید. به این ترتیب و با توجه به اینکه در زبان فارسی سه ساختار هجایی "CV"، "VC" و "CVCC" وجود دارند، مجموعا ۳ درخت تصمیم تعمیم‌یافته برای کلمات تک هجایی، ۹ درخت برای کلمات دو هجایی و ۲۷ درخت برای کلمات سه هجایی می‌بایست در نظر گرفت.




شکل ۲- فرآیند تولید گونه‌های تلفظی کلمات

درخت تصمیم تعمیم یافتهٔ ایستا، برای پیش‌بینی اینکه کدام واج‌ها در کلمه می‌توانند تبدیل یا حذف شوند و یا در کدام قسمت‌ها درج واج می‌تواند اتفاق بیافتد، خصوصیات همخوان‌ها و واکه‌ها در نواحی مختلف کلمه، همچنین موضع هجای تکیه دار را به عنوان ورودی دریافت می‌کند. درخت‌های پویا علاوه بر اینها به نرخ گفتار نیز توجه می‌نمایند. خصوصیات همخوان‌ها و واکه‌ها برای این درخت‌ها، براساس تعلق آنها به گروه‌های مختلف واجی بیان می‌شود. دسته‌بندی واج‌ها به گروه‌های مختلف براساس تشابه واج‌ها از نظر زبانشناسی و یا ماتریس ابهام واج‌ها صورت پذیرفته است. منظور از الگوی تغییرات که در شکل بیان شده است، الگویی است که گونه تلفظی براساس آن از گونه مرجع تولید می‌شود. هر الگوی تغییرات نشان می‌دهد کدامیک از واج‌های کلمه قابل تبدیل و یا حذف می‌باشند و یا در کجای کلمه درج واج می‌تواند رخ دهد. توسط درخت، به هر الگوی تغییرات احتمالی تخصیص می‌یابد. برای تولید گونه‌های تلفظی، حدود آستانه‌ای جهت تعیین "احتمال قابل قبول" تعیین می‌شود تا تعداد الگوهای تغییرات پیش‌بینی شده توسط درخت، محدود شوند. در مرحله بعدی قواعد تلفظی به آن نواحی که توسط درخت مجاز به تغییر شناخته شده اند، اعمال می‌شوند. در این مرحله، قواعد تغییر واج‌ها، به مواضع تعیین شده در کلمه اعمال می‌شوند. در بخش ۶ به تعریف این قواعد تلفظی می‌پردازیم. با اعمال قواعد تلفظی، گونه‌های تلفظی کلمات تولید می‌شوند. بعد از تولید گونه‌ها، درست نمایی[۲۵] آنها نرمالیزه می‌شوند تا مجموع احتمالات وقوع آنها مساوی یک شود. برای آزمایش کیفیت گونه‌های تلفظی تولید شده، می‌توان آنها را در واژگان یک سیستم بازشناسی گفتار پیوسته فارسی قرار داد و سپس اثر آنها را در افزایش درصد صحت بازشناسی ارزیابی نمود. چنانچه از مدل‌های پویا برای تولید گونه‌های تلفظی استفاده شود، گونه‌های موجود در واژگان سیستم، با تغییرات نرخ گفتار ورودی، بصورت پویا تغییر می‌یابند.

## ۳- دادگان مورد استفاده

دادگان مورد استفاده در این تحقیق، بخشی از "فارس‌دات بزرگ" بوده است. فارس‌دات بزرگ، یک دادگان گفتاری فارسی است که توسط مرکز تحقیقات پردازش هوشمند علائم تولید شده است. این دادگان در برگیرنده گفتار ۱۰۰ گوینده است. در انتخاب گوینده‌ها سعی شده است تنوعات کافی گوینده‌ها شامل تنوعات سن، جنسیت، سطح تحصیلات و تعلق آنها به لهجه‌های ده‌گانه فارسی وجود داشته باشد. لهجه‌های موجود در ایران شامل تهرانی، ترکی، اصفهانی، جنوبی، شمالی، خراسانی، بلوچی، لری و یزدی می‌باشند. هر گوینده در حدود ۴۰۰۰ کلمه از متنهای گوناگون روزنامه‌ها را در محیط یک اتاق اداری بیان نموده است. متون خوانده شده، زمینه‌های گوناگونی شامل سیاسی، اقتصادی، فرهنگی، ورزشی و ... را شامل می‌شوند. برچسب‌دهی واجی جملات با استفاده از کاراکترهای IPA با شکلی مشابه فارس‌دات [۲۸] انجام گرفته است. برای تولید دادگان مناسب برای آموزش مدل‌های تلفظی، ابتدا با استفاده از موتور بازشناسی واج، برداشته شده از سیستم بازشناسی گفتار پیوسته "شنوا"، بازشناسی دنباله آواهای بخش بزرگی از سیگنال جملات فارس‌دات بزرگ انجام شد[۲۹]. در قدم بعدی دنباله واج‌های بازشناسی شده (نسخه‌های تولید شده بصورت خودکار توسط بازشناس واج) بصورت خودکار با نسخه‌های واجی مرجع همردیف‌سازی گردیدند. همردیف سازی خودکار با استفاده از یک الگوریتم برنامه‌ریزی پویا که سعی بر کمینه کردن فاصله دنباله‌های همردیف‌سازی شده دارد، انجام می‌گیرد. هزینه همردیف‌سازی دراین الگوریتم براساس اختلاف ویژگی‌های بین واج‌های متناظر در نسخه مرجع و نسخه بازشناسی شده محاسبه می‌گردد، یک مثال از همردیف‌سازی برای کلمه "کتابخانه" با نسخه واجی مرجع /**ketabxaneh**/ ارائه می‌شود.

/**ketabxaneh**/
/**Peta#fane#**/

ردیف اول در این مثال نسخه واجی مرجع کلمه و ردیف دوم دنباله واجی بازشناسی شده می‌باشد. به تبدیل واج انفجاری بی‌صدا **ə pə** به واج انفجاری بی‌صدا **k** (**k→p**)، حذف واج انفجاری صدادار (**#→b**)، تبدیل همخوان (**x→f**) و حذف واج سایشی بی‌صدا (**#→h**) توجه نمایید. نماد "#" برای نمایش حذف و درج واجها استفاده شده است. همانطور که اشاره شد در این تحقیق دنباله‌های بازشناسی شده با دنباله‌های مرجع کلمات همردیف‌سازی می‌شوند. تفاوت بین این دو دنباله ناشی از دو عامل است. عامل اول گوینده است که لزوما کلمات را بصورت مرجع تلفظ نمی‌نماید، و عامل دوم سیستم بازشناس واج است که مسلما عاری از خطا نیست، لذا ممکن است در بسیاری از موارد واج‌های بیان‌شده را به اشتباه بازشناسی کند. در برخی از روش‌ها دنباله‌های بیان شده توسط گویندگان را بصورت دستی برچسب می‌زنند و آنها را با دنباله‌های مرجع همردیف‌سازی می‌کنند تا مدل‌های تلفظی را آموزش دهند. اما اینجا مدل‌سازی همزمان تنوعات ناشی از گوینده و ناشی از خطاهای سیستم بازشناس واج مورد توجه است. اگرچه متغیرهای متفاوتی این دو منبع بروز تنوعات واجی را تحت تاثیر قرار می‌دهند، اما ما بر این اساس عمل نموده‌ایم که مدل‌های آماری طراحی شده، قادر به مدلسازی همزمان تنوعات ناشی از هر دو منبع هستند.

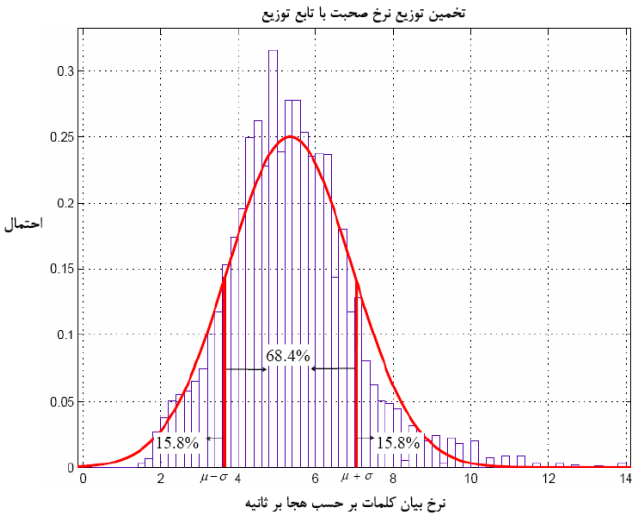
## ۴- تحلیل تاثیر عوامل گوناگون بر بروز تنوعات تلفظی واجی

اهمیت ساختار واجی-هجایی کلمه، زمینه متنی واجی و موضع تکیه در کلمه در بروز تنوعات تلفظی که شامل تبدیل، حذف و درج واج می‌شوند، روشن است، و در منابع زبانشناسی بسیاری مورد تاکید قرار گرفته است [۳۰ و ۳۱]. محققان بسیاری از این ویژگی‌ها در مدلسازی تنوعات تلفظی استفاده نموده‌اند [۱۰، ۱۱، ۱۳، ۲۱ و ۲۳]. قرار گیری واجی خاص در یک متن واجی بخصوص، می‌تواند باعث حذف یا تبدیل آن در حین تلفظ شود. همچنین هجاهای تکیه‌دار در کلمه معمولا با شدت بیشتر و بصورت شمرده‌تری بیان می‌شوند و این باعث می‌شود این هجاها کمتر دچار تغییرات واجی شوند. نرخ صحبت و قابلیت پیش‌بینی کلمه نیز از عواملی هستند که نقش مهمی در بروز این تغییرات دارند. در ادامهٔ این گزارش برآنیم که قدری به تحلیل کمی‌تری از میزان و نحوه تاثیر دو ویژگی اخیر، یعنی نرخ صحبت و قابلیت پیش‌بینی کلمه بر بروز تنوعات تلفظی واجی بپردازیم. در انتهای این بخش به بررسی توزیع فراوانی مشترک دو پارامتر نرخ صحبت و احتمال یک تایی برای واحد های کلمه خواهیم پرداخت و تاثیر مشترک و همزمان این دو پارامتر را روی میزان انحراف تلفظ واقعی از تلفظ مرجع بررسی خواهیم نمود.

### ۴-۱- تحلیل آماری تاثیر نرخ صحبت بر تنوعات تلفظی واجی

معمولاً نرخ صحبت بر اساس تعداد واحدهای زبانی بیان شده در هر ثانیه اندازه‌گیری می‌شود. در این گزارش، از واحد تعداد هجا بر ثانیه برای اندازه‌گیری نرخ صحبت استفاده می‌شود. همانطور که در دیگر منابع بررسی شده است در نرخ بالای گفتار، دنباله‌های واجی بیان شده توسط گوینده‌ها کوتاهتر از دنباله‌های واجی مرجع کلمات می‌باشند [۲۲]. تابع توزیع احتمال نرخ صحبت برای تمام کلمات موجود در دادگان "فارس‌دات بزرگ" برحسب هجا بر ثانیه در شکل ۳ آمده است. این تابع توزیع به تابع گوسی نزدیک

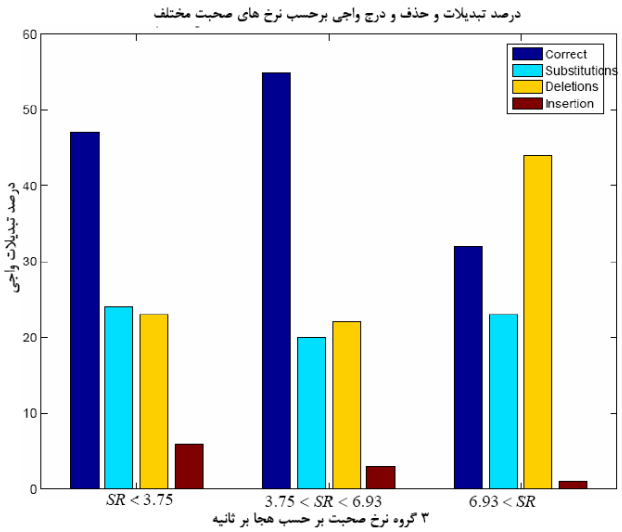
است و با تقریب خوبی، یک تابع توزیع احتمالات گوسی بر آن تطبیق داده شده است. میانگین نرخ صحبت کلمات در کل دادگان ۵/۳۴ هجا بر ثانیه و انحراف معیار آن ۱/۵۹ هجا بر ثانیه می‌باشد. دادگان را بر اساس میانگین و انحراف معیار به سه قسمت تقسیم نموده‌ایم. این سه قسمت شامل نواحی "نرخ صحبت کم"، "نرخ صحبت متوسط" و "نرخ صحبت زیاد" می‌باشند.




شکل ۳- تابع توزیع احتمال نرخ صحبت کلمات بر حسب هجا بر ثانیه و تقسیم بندی دادگان به سه بخش "نرخ صحبت کم"، "نرخ صحبت متوسط" و "نرخ صحبت زیاد" بر اساس میانگین و انحراف معیار نرخ صحبت

همانطور که در شکل ملاحظه می‌شود قسمت عمده دادگان (۶۸/۴٪ از کل دادگان) در بخش نرخ صحبت متوسط قرار گرفته است. با توجه به اینکه دادگان "فارس‌دات بزرگ" بصورت گفتار روخوانی است و معمولا گوینده‌ها در این نوع گفتار با نرخ صحبت متوسط به صحبت می‌پردازند؛ فرض بر این است که دادگان فارس‌دات حاوی مقدار کمی گفتار سریع و گفتار آهسته است. بررسی‌ها نشان می‌دهد که این مقدار هم بیشتر بدلیل خصوصیات ذاتی گوینده‌ها است، چراکه برخی از گوینده‌ها بطور ذاتی تند و یا کند صحبت می‌کنند. لذا به منظور تطبیق بیشتر مرزبندی نرخ صحبت با طبیعت توزیع نرخ صحبت در فارس‌دات، قسمت عمده دادگان را در دسته نرخ صحبت متوسط قرار دادیم. البته در آموزش درخت‌های تصمیم، نرخ صحبت بعنوان یک ویژگی پیوسته و نه گسسته به درخت داده می‌شود، و در آموزش درخت‌های پویا، نرخ صحبت بعنوان یک ویژگی گسسته از انواع کم، متوسط یا زیاد مورد استفاده قرار نمی‌گیرد. این بحث مرزبندی نرخ صحبت (به صورت شکل ۳) هم که در این جا مطرح نموده‌ایم، تنها برای نشان دادن ارتباط کمّی نرخ‌های صحبت متفاوت (در ۳ دسته)، با تبدیلات واجی است که در کلمات آدا شده رخ می‌دهند.

بعد از آنکه دادگان به سه بخش فوق تقسیم گردید، دنباله‌های واجی مرجع و بازشناسی شده کلمات، در هر بخش از دادگان همردیف‌سازی شدند. سپس با توجه به دنباله‌های همردیف‌شده، آماری از میزان تطابق، تبدیل، حذف و درج واجها در هر بخش از دادگان گفتاری تهیه گردید. نتایج در شکل ۴ آورده شده‌اند. در این نمودار درصد تبدیلات واجی برای هر یک از بخش‌های دادگان ارائه گردیده است. رنگ آبی تیره بیانگر درصد تطابق واج، رنگ آبی روشن بیانگر درصد تبدیل واج، رنگ زرد بیانگر درصد حذف واج و رنگ قرمز بیانگر درصد درج واج می‌باشد. شکل ۴ درصد تطابق، تبدیل، حذف و درج واجها را براساس تغییر نرخ صحبت در کل دادگان نشان می‌دهد.




شکل ۴- درصد تبدیل، حذف و درج واجها برای گروه های سرعتی مختلف گفتار

همانطور که در نمودار ملاحظه می‌شود، در نرخ گفتار متوسط (بین ۳/۷۵ تا ۶/۹۳ هجا بر ثانیه)، بیشترین تطابق بین دنباله‌های واجی بازشناسی‌شده و مرجع وجود داشته است. از طرفی در این نرخ گفتار، دنباله‌های واجی کلمات با کمترین میزان حذف واج نسبت به دنباله واجی مرجع کلمات توسط گوینده‌ها ادا می‌شوند. البته در این نرخ گفتار، سیستم بازشناسی گفتار "شنوا" نیز حالت بهینه دارد. با افزایش نرخ صحبت در نرخ‌های بالاتر از ۶/۹۳ هجا بر ثانیه، میزان حذف واج به شدت افزایش می‌یابد. این بدان معناست که با افزایش نرخ صحبت، تحت تأثیر دو پدیده استفاده از دنباله‌های واجی کوتاه توسط گوینده و نیز کاهش کارآیی سیستم بازشناسی آوا، بسیاری از واجها نسبت به دنباله واجی مرجع کلمه حذف می‌گردند. همانطور که در نمودار مشخص است در نرخ‌های صحبت کمتر از ۳/۷۵ هجا بر ثانیه، کارآیی سیستم از حالت بهینه دور می‌شود. تحلیل این نمودار، ضرورت مدلسازی و وارد نمودن عامل "نرخ صحبت" را در تهیه یک واژگان پویا در سیستم بازشناسی گفتار آشکار می‌سازد. منظور از واژگان پویا، واژگانی است که گونه‌های تلفظی کلمات موجود در آن، با توجه به تغییر نرخ صحبت گفتار ورودی تغییر می‌کنند، به نحوی که مناسب برای نرخ صحبت موجود باشند. از ویژگی‌های چنین واژگانی این است که، در نرخ‌های صحبت بالا به گونه‌های تلفظی با طول واجی کوتاهتر احتمال بیشتری را اختصاص خواهد داد. همچنین، در نرخ صحبت متوسط، تعداد گونه‌های تلفظی مورد استفاده برای هر کلمه کاهش می‌یابد، چرا که در این نرخ صحبت معمولاً کلمات بسیار نزدیک‌تر به همان صورت دنباله واجی مرجع تلفظ می‌شوند. همچنین بهتر است احتمال گونه مرجع کلمات در این نرخ صحبت افزایش پیدا کند. مدل‌سازی عامل نرخ صحبت در سطح واژگان سیستم می‌تواند باعث کاهش خطای بازشناسی کلمات گردد. در بسیاری از منابع نشان داده شده است که خطای سیستم‌های بازشناسی گفتار، تحت تأثیر نرخ صحبت افزایش می‌یابد [۲۲].

در تحلیل دیگری از معیار آنتروپی برای نشان دادن تأثیر نرخ صحبت بر تنوعات تلفظی واجی استفاده نمودیم. آنتروپی می‌تواند معیاری از میزان تنوعات تلفظی در سطح دادگان ارائه دهد. اگر یک آوائی (برای مثال واج)، در سطح دادگان به صورت‌های گوناگونی بازشناسی شده باشد، و یا میزان حذف و تبدیل آن زیاد باشد، آنتروپی آن بالا خواهد بود. و به عکس اگر در موارد کمی، تبدیل یا حذف شده باشد، آنتروپی آن کم خواهد بود. آنتروپی تلفظی یک واحد زبانی نظیر واج، اندازه‌ای از تنوعات تلفظی آن واحد

زبانی در دادگان مورد بررسی می‌باشد. فرض کنیم واج $\alpha$ بصورت واجهای موجود در مجموعه $X$ قابل تلفظ باشد؛ در این صورت آنتروپی تلفظی واج $\alpha$ را به صورت زیر تعریف می‌نمائیم:

$$H(\alpha) = \sum_{x \in X} p(x) \log p(x) \qquad (۲)$$

به طوریکه $p(x)$ احتمال تلفظ $\alpha$ بصورت $x$ می‌باشد. درصورتیکه تمام تلفظهای موجود در مجموعه $X$ احتمال برابر داشته باشند، آنتروپی بیشینه خواهد بود و وقتی $\alpha$ فقط به یک صورت تلفظ گردد، آنتروپی کمینه با مقدار صفر حاصل خواهد شد. برای مثال اگر احتمالات بازشناسی واج ɔ در کل دادگان را به صورتهای ɔ(تطابق صحیح)، **a**، **e** و **o**، به ترتیب با $p(\partial)$ ، $p(a)$ ، $p(e)$ و $p(o)$ نشان دهیم، آنتروپی این واج بصورت زیر محاسبه خواهد شد:

$$H(\partial) = \sum_{x \in \{\partial, a, e, o\}} p(x) \log p(x) \qquad (۳)$$

$p(x)$ بیانگر احتمال تلفظ ɔ بصورت $x$ می‌باشد. و بصورت زیر محاسبه می‌شود:

$$p(x) = \frac{N(\partial \to x)}{N(\partial)} \qquad (۴)$$

$N(\partial \to x)$ بیانگر تعداد مواردی است که واج ɔ بصورت $x$ بازشناسی شده است، و $N(\partial)$ تعداد کل واجهای ɔ در نسخه‌های واجی مرجع است. جدول ۱، میانگین آنتروپی واجها را که بصورت میانگین آنتروپی‌های محاسبه شده برای واجها روی ۲۹ واج زبان فارسی محاسبه شده‌اند را برای دسته‌های سه‌گانه نرخ صحبت نشان می‌دهد. همانطور که دیده می‌شود، آنتروپی به عنوان معیاری از تنوعات تلفظی با تغییر نرخ صحبت تغییر می‌کند. افزایش میانگین آنتروپی واجی به معنای افزایش اختلاف بین واجهای متناظر در نسخه‌های واجی مرجع و بازشناسی شده کلمات است. این بدان معناست که گوینده واجها را به صورتهایی غیر از صورت‌های مرجع تلفظ می‌کند و/یا واجها با کاهش کارآیی سیستم بازشناس واج بصورت واجهای دیگر بازشناسی می‌شوند. میانگین آنتروپی واجی برای نرخ صحبت بین ۳/۷۵ تا ۶/۹۳ هجا بر ثانیه کمترین مقدار را دارد. این نکته نشان می‌دهد که در این نرخ صحبت، بیشترین تطابق بین نسخه‌های واجی مرجع و بازشناسی شده کلمات وجود داشته است. به عبارتی، تلفظ گوینده به تلفظ مرجع نزدیکتر بوده و کارآیی سیستم بازشناسی واج نیز دراین نرخ صحبت بیشتر بوده است.

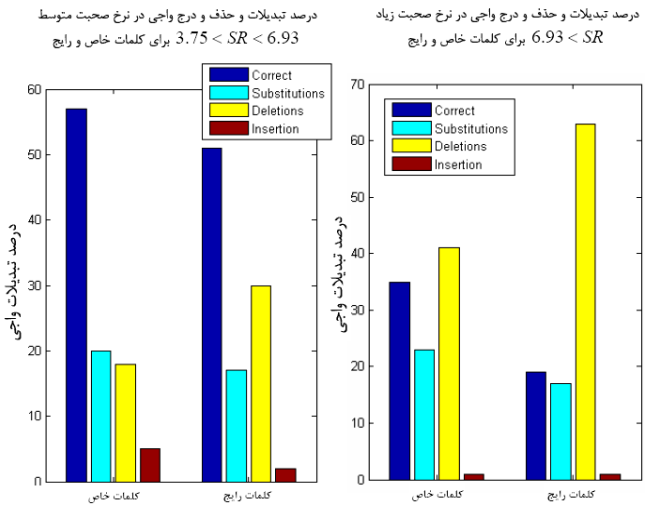
جدول ۱- میانگین آنتروپی واجی برای سرعت‌های مختلف گفتار

| **نرخ گفتار** | کمتر از ۳/۷۵ | بین ۳/۷۵ و ۶/۹۳ | بیشتر از ۶/۹۳ |
|---|---|---|---|
| **آنتروپی** | ۱/۶۱ | ۱/۵۴ | ۱/۶۶ |

در نرخهای صحبت کم و زیاد، آنتروپی سیستم تحت تاثیر افزایش تنوعات در تلفظ گوینده و نیز کاهش کارآیی سیستم بازشناسی آوا افزایش می‌یابد. با مدلسازی تاثیر نرخ صحبت در سطح واژگان سیستم می‌توان با انتخاب گونه‌های مناسب از آوانویسی کلمات در واژگان سیستم بازشناسی، به نرخ صحت بازشناسی بالاتری دست یافت. این کار از طریق اندازه‌گیری نرخ صحبت گفتار ورودی و بکارگیری آن در اصلاح گونه‌های کلمات موجود در واژگان، به صورتی پویا و تطبیقی نسبت به تغییر لحظه‌ای سرعت آدای کلمات دست یافتنی است.

## ۴-۲- تحلیل آماری تاثیر احتمال وقوع کلمه بر تنوعات تلفظی واجی

اثبات شده است که احتمال وقوع کلمه بر نحوه تلفظ گوینده تاثیر می‌گذارد [۶، ۲۳]. این تاثیر با مبانی تئوری اطلاعات قابل توجیه است. هر چه احتمال بیان یک کلمه بیشتر باشد، با توجه به اینکه شنونده بیشتر انتظار شنیدن آن را دارد، خیلی راحت‌تر آنرا بازشناسی می‌نماید. در چنین مواردی گوینده بطور ناخودآگاه از گونه‌های تلفظی با طول واجی کمتر استفاده می‌نماید، و معمولا آنها را سریعتر آدا می‌نماید. اما در مواردی که کلمات خاص و یاغیر منتظره آدا می‌شوند، دقت بیشتری در تولید آنها می‌شود، و معمولا آهسته‌تر بیان می‌شوند. در چنین مواردی گوینده از گونه‌های تلفظی رایج‌تر استفاده می‌نماید، تا به شنونده در بازشناسی آنها کمک شود. تاثیر احتمال وقوع کلمات بر تلفظ کلمات، یک الگوی ذاتی در گفتار تمام انسانها است [۲۳]. از منظر دیگر، کلمات خاص معمولا همان کلماتی در جمله هستند که حاوی اطلاعات و معنا می‌باشند. چنین کلماتی بطور طبیعی با دقت بیشتر و با گونه واجی نزدیک‌تر به گونهٔ مرجع ادا می‌شوند. به عکس کلماتی که فاقد اطلاعات باشند، به صورتهای تقلیل یافته و با سرعت زیاد آدا می‌شوند. در این قسمت برای تحلیل آماری تاثیر احتمال وقوع کلمه بر تنوعات تلفظی از احتمالات یک-تائی (Unigram) کلمات، به عنوان معیاری از احتمال وقوع کلمات استفاده شده است. برای نشان دادن کمّی اثر این عامل، کلمات درون دادگان را به دو بخش تقسیم نمودیم. قسمت اول شامل کلمات رایج با تعداد تکرار بیشتر از ۱۰۰ مورد در کل دادگان و قسمت دوم شامل کلمات خاص با تعداد تکرار کمتر از ۱۰۰ مورد می‌شود. سطح آستانهٔ ۱۰۰، با توجه به حجم دادگان و تقسیم ان به دو بخش نسبتاً متعادل، انتخاب گردید. سپس درصد تبدیلات واجی برای هر قسمت محاسبه شد. از آنجاییکه نرخ گفتار نیز یک عامل تاثیرگذار در تبدیلات واجی است، در حین مطالعه اثر احتمال وقوع کلمات در تنوعات تلفظی، بهتر است نرخ گفتار را ثابت یا حداقل در محدوده ثابتی درنظر گرفت. لذا درصد تبدیلات واجی برای کلمات رایج و خاص در دو حالت سرعت متوسط و زیاد (برای مثال) ارزیابی شده و نتایج در شکل ۵ آورده شده‌اند.

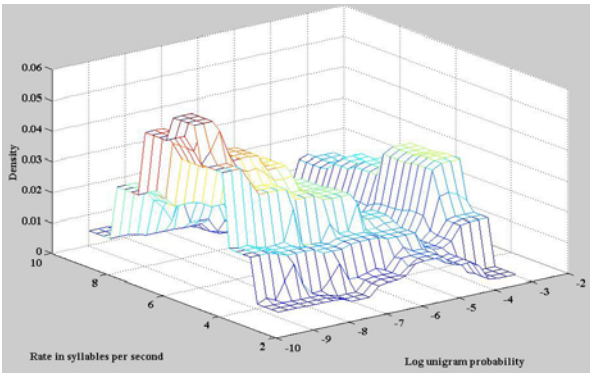



شکل ۵- درصد تبدیل، حذف و درج واجها برای کلمات خاص و رایج در دو دسته نمونه نرخ گفتار متوسط و زیاد

همانطور که در نمودارها ملاحظه می‌شود، درصد تطابق و تبدیل واج در کلمات رایج، نسبت به کلمات خاص کمتر است؛ برعکس، درصد حذف واج در کلمات رایج بیشتر از کلمات خاص است. با توجه به نتایج می‌توان گفت، با افزایش احتمال وقوع کلمه، گوینده از گونه‌های تلفظی با دنباله‌های واجی کوتاه‌تر استفاده می‌نماید. توجه به این نکته لازم است که کارآیی سیستم بازشناسی واج مستقیما از احتمال وقوع کلمه تاثیر نمی‌پذیرد، و برای کلمات رایج و خاص یکسان است. بنابراین بهتر است برای کلمات رایج از گونه‌های تلفظی با طول واجی کوتاهتر در واژگان استفاده شود. احتمال تلفظ کلمات رایج بصورت مرجع بسیار کمتر از این مقدار برای کلمات خاص است. با مدل‌سازی تاثیر احتمال وقوع کلمه بر تنوعات تلفظی واجی، می‌توان به بهینه‌سازی احتمالات تخصیص یافته به گونه‌های تلفظی کلمات در سیستم بازشناسی گفتار پیوسته پرداخت. مدل‌های ترکیبی ایستا و پویا در فرآیند تولید گونه‌های تلفظی کلمات از احتمال "یک- تائی" بعنوان یک ویژگی ورودی استفاده می‌نمایند. این نوع از احتمالات، قبلا از یک دادگان متنی بصورت جداگانه استخراج شده‌اند و در این کار مورد استفاده قرار می‌گیرند.

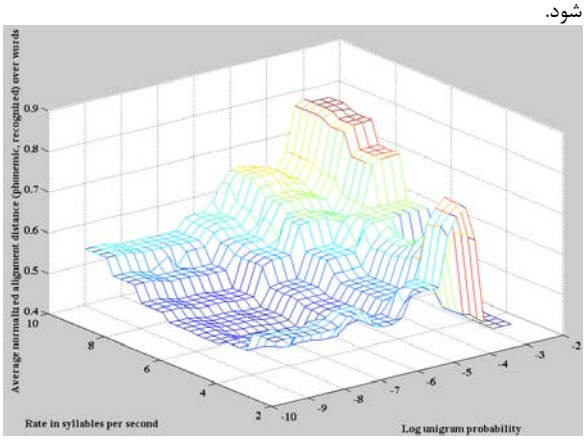
## ۴-۳- بررسی همزمان تاثیر دو پارامتر نرخ صحبت و احتمال یک تایی روی انحراف تلفظ واقعی از تلفظ مرجع در واحد های کلمه

اگر دو پارامتر نرخ صحبت و احتمال یک تایی را بعنوان دو متغیر در نظر بگیریم با بررسی نمودار توزیع فراوانی مشترک آنها می توان دریافت که دو متغیر مستقل از یکدیگر نیستند. شکل ۶ توزیع فراوانی مشترک نرخ صحبت و لگاریتم احتمال یک-تایی را نشان می دهد. طبعا مشخص است که کلمات رایج دارای لگاریتم احتمال یک-تایی بالاتری می باشند درحالیکه کلمات خاص دارای لگاریتم احتمال یک تایی پایینتری می باشند. همانطور که در شکل ۶ مشخص است برای کلمات رایج فراوانی نرخ صحبت در نرخ های بالا بسیار بیشتر از نرخ های کمتر از ۴ هجا بر ثانیه است ؛ حال آنکه برای کلمات رایج فراوانی در نرخ های متوسط متمرکز است و در نرخ های پایین و بالا کم می باشد. بنابراین دو فوق روی یکدیگر تاثیر متقابل دارند. و معمولا کلماتی که قابلیت پیشبینی زیادی دارند با نرخ بیشتری بیان می گردند.




شکل ۶- توزیع فراوانی مشترک نرخ صحبت و لگاریتم احتمال یک تایی در واحدهای کلمه استخراج شده از کلمات موجود در دادگان فارس دات بزرگ

حال که مشخص شد دو پارامتر نرخ صحبت و قابلیت پیش بینی کلمه در توزیع فراوانی مستقل نیستند. تاثیر همزمان آنها را روی تنوعات تلفظی بررسی می نماییم تا ببینیم آیا تاثیر آنها بر تلفظ یک تاثیر مستقل و جمع آثار است یا اینکه یک اثر پیچیده و متقابل است. شکل ۷ میزان انحراف تلفظ های واقعی را از تلفظ مرجع در مقادیر مختلف نرخ صحبت و لگاریتم احتمال یک تایی نشان می دهد. میزان انحراف تلفظ های واقعی از تلفظ های مرجع در محور عمودی با محاسبه میانگین فواصل همردیف سازی (بین گونه واقعی و مرجع) نرمالیزه شده به طول واجها در کلمات نشان داده شده است. همانطور که ملاحظه می شود بیشترین انحراف از تلفظ های مرجع در کلمات رایج و نرخ های صحبت بالا رخ می دهد. و کمترین تغییرات تلفظی برای کلمات خاصی که در نرخ های متوسط بیان می شوند دیده می شود.




شکل ۷- میزان انحراف تلفظ های واقعی از تلفظ های مرجع در نرخ های صحبت و احتمال یک تایی های مختلف

اثر متقابل پیچیده ای در تاثیر گذاری این دو متغیر روی تنوعات تلفظی دیده نمی شود و می توان برخی تاثیرات غیرخطی قابل مشاهده در شکل را به حساب تعداد ناکافی نمونه ها در برخی نواحی دانست. بنابراین تا این لحظه این فرض ساده سازی قابل قبول است که می توان تاثیر همزمان این دو پارامتر را بر تنوعات تلفظی مستقل دانست و اثر همزمان آنها را بصورت یک تاثیر جمع آثار پذیرفت.

## ۵- درخت تصمیم تعمیم یافته و ویژگی‌های ورودی

به هر گروه از کلمات با ساختار هجایی مشابه، یک درخت تصمیم تعمیم یافته تخصیص و آموزش می‌یابد. کلمات براساس تعداد و ساختار واجی هجاها گروه‌بندی می‌شوند. در فارسی سه ساختار هجایی داریم که شامل "CV" ، "CVC" و "CVCC" می‌باشند. "C" بیانگر همخوان و "V" بیانگر واکه می‌باشد. براین اساس، برای مدل‌سازی تنوعات تلفظی به روشی که پیشنهاد نموده‌ایم، در زبان فارسی ۳ گروه کلمات تک هجایی، ۹ گروه مختلف کلمات دو هجایی و ۲۷ گروه کلمات سه هجایی باید بطور جداگانه مدلسازی شوند. اصطلاح درخت تعمیم یافته، نام خود را از این ایده می‌گیرد که آنگونه که در [۲۲ و۲۳] دیده می‌شود، برای هر کلمه یک درخت تخصیص نیافته است بلکه هر درخت مختص یک گروه از کلمات با ساختار هجایی یکسان است. مثلا گونه‌های تلفظی دیده شده برای دو کلمه "کتاب" با دنباله واجی مرجع /***ketab***/ و "مداد" با دنباله واجی مرجع /***medad***/ هر دو برای آموزش درخت "CVCVC"استفاده می‌شوند. درخت‌های تصمیم برای آموزش، به ویژگی‌های واج‌های مختلف کلمه و نیز موضع هجای تکیه‌دار در کلمه توجه می‌نمایند. درخت‌های تصمیم در مدل‌های پویا علاوه بر اینها به نرخ صحبت گوینده نیز در فرآیند تولید گونه‌های تلفظی توجه می‌نمایند. واج‌ها بر اساس تعلق آن‌ها به دسته‌های واجی، که بر اساس شباهت آوائی واج‌ها تشکیل یافته‌اند، مورد تصمیم‌گیری درخت ها قرار می‌گیرند. جدول ۲، تعداد ۷ دسته از همخوان‌ها را به همراه ویژگی آوائی اصلی هر دسته ارائه می‌دهد. واکه‌ها هم با توجه به ماتریس ابهام واج استخراج شده از دادگان (احتمال تبدیل به یکدیگر در حین بازشناسی توسط سیستم بازشناسی گفتار)، به سه دسته $\{\mathbf{e}, \mathbf{i}\}$ ، $\{\mathbf{ɔ}, \mathbf{a}, \mathbf{o}\}$ و $\{\mathbf{u}\}$ تقسیم شده‌اند.

جدول ۲- گروه‌های هفت گانه همخوان‌ها و توصیف زبانشناسی آنها

| ویژگی از نظر زبانشناسی | نماد IPA |
|---|---|
| انفجاری صدادار | $\mathbf{b, d, G, ʤ, g}$ |
| انفجاری بی‌صدا | $\mathbf{p, t, k, ʧ}$ |
| سایشی بی‌صدا | $\mathbf{s, ʃ, x, f}$ |
| همخوان رسا | $\mathbf{l, m, n, r}$ |
| سایشی صدادار | $\mathbf{z, ʒ, v}$ |
| همخوان حلقی | $\mathbf{ʔ, h}$ |
| روان | $\mathbf{j}$ |

الگوریتم آموزش درخت‌ها سعی می‌کند به کمترین میزان آنتروپی در گره‌های نهایی درخت دست یابد. تعداد زیاد الگوهای تبدیلات واجی بین نسخه‌های واجی مرجع و بازشناسی‌شده کلمات(موجود در دادگان مورد استفاده) و تعداد تکرار محدود برخی از آنها، یک چالش در آموزش درخت‌های تصمیم تعمیم یافته می‌باشد. برای رفع این مشکل با استفاده از یک تکنیک کوانتیزاسیون برداری، به کوانتیزه نمودن بردار کد متناظر با هر الگوی تبدیلات واجی برای هر داده پرداختیم، تا تعداد زیاد الگوهای مشاهده شده را کاهش دهیم. فرآیند کوانیتزاسیون نزدیکترین سطح برداری کوانتیزاسیون را با محاسبه فاصله سطوح برداری، با بردار کد متناظر الگوی تبدیلات پیدا می‌کند. سپس اندیس این سطح برداری، به عنوان کدی برای الگوی تبدیلات پردازش شده مورد استفاده قرار می‌گیرد. استفاده از این رویکرد، تعداد الگوهای تبدیلات مختلف دیده شده برای هر داده (جفت نسخه‌های واجی مرجع و بازشناسی شده) را کاهش

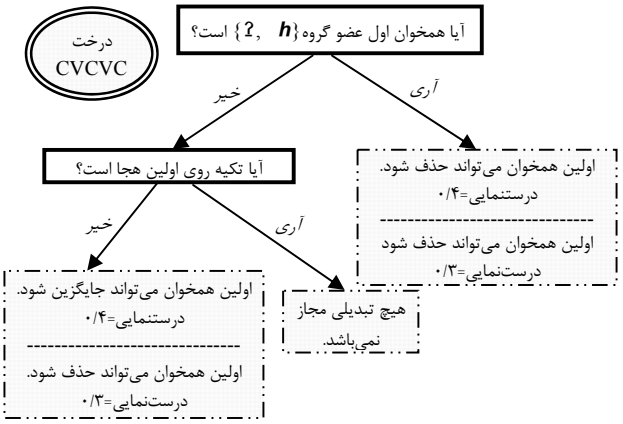
می‌دهد و به ازای آن، تعداد تکرار الگوهای تبدیلات واجی بعد از کوانتیزاسیون افزایش میدهد. لازم به ذکر است، که سطوح برداری کوانتیزاسیون، با خوشه‌بندی برداری الگوهای تبدیلات (با استفاده از الگوریتم "k-means") انتخاب می‌شوند. بردارهای کد متناظر با الگوهای تبدیلات واجی، در ابعاد $4N+1$ هستند (N طول واجی کلمات است). این بردارها شامل عناصر ۰ و ۱ می‌باشند. هر عنصر یک بردار نمایانگر وقوع/عدم وقوع حالت‌های تطابق، تبدیل، حذف یا درج واجها می‌باشد. برای هر واج ۳ عنصر مجزا در بردار کد در نظر گرفته شده است. عنصر اول متناظر با تطبیق (بازشناسی شده به همان صورت نسخه مرجع)، عنصر دوم متناظر با تبدیل واج و عنصر سوم متناظر با حذف واج می‌باشد. درج واجها با تخصیص یک عنصر میانی با مقدار ۱ در بین سه عنصر متناظر با دو واج مجاور و همچنین در ابتدا و انتهای بردار کد بیان می‌شوند (به این ترتیب تعداد کل عناصر بردار متناظر با کلمه برابر $4N+1$ خواهد شد). فرض کنید یک کلمه با ساختار واجی "CVCVC" بصورت "C'VCV#" (تبدیل اولین همخوان و حذف آخرین همخوان) بازشناسی شده باشد. بنا به آنچه گفته شد بردار کد متناظر با الگوی تبدیل این داده بصورت بردار زیر با طول $4\times5+1$ می‌باشد.

0 0 1 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0

*i m s d i m s d i m s d i m s d i m s d i*

"i" بیانگر عناصر درج واج، "m" بیانگر عناصر تطبیق واج، "s" بیانگر عناصر تبدیل واج و "d" بیانگر عناصر حذف واج در بردار کد می‌باشند. این عناصر با توجه به وضعیتی که برای واج رخ داده است برابر ۱ یا ۰ قرار داده می‌شوند. عناصر متناظر با درج واج نیز اگر درج واجی بین دو واج مجاور رخ داده باشد، با ۱ در موضع متناظر مشخص می‌شوند.

شکل ۸ بخشی از درخت تصمیم تعمیم‌یافته را با ساختار "CVCVC"، بعد از کوانتیزاسیون برداری الگوهای تغییرات و انجام مرحلهٔ آموزش، نشان می‌دهد.



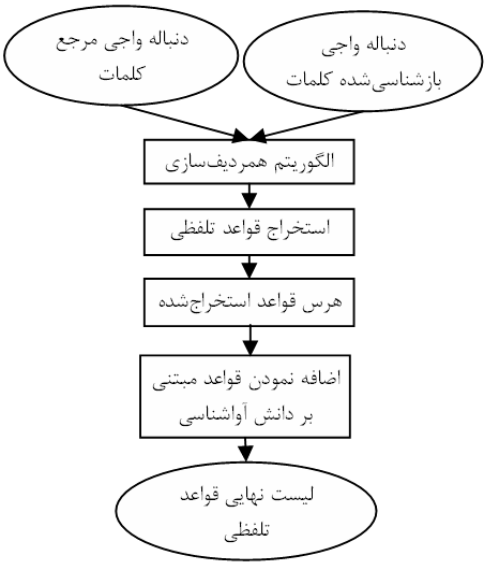

شکل ۸- مثالی از درخت تصمیم تعمیم یافته "CVCVC"

از آنجائیکه کلمات با تعداد بیش از چهار هجا بسیار نادر هستند، نمی‌توان برای آنها مدلهای مناسبی را آموزش داد. لذا برای تولید گونه‌های تلفظی این کلمات فقط از قواعد تلفظی استفاده نمودیم.

همانطور که قبلاً هم بیان شد هر درخت تصمیم برای گروه کلمات با ساختار هجایی مشابه آموزش می‌یابد. در حین فرآیند آموزش، آلگوریتم آموزش در مورد خصوصیات واج‌ها، موضع هجای استرس دار در کلمات پرسش می‌کند تا بتواند دادگان را در گره‌های جدید خود تفکیک نماید. بنابراین در گره‌های نهایی درخت تصمیم، کلماتی با واج‌های مشابه از نظر زبانشناسی در مواضع متناظر جای می‌گیرند. پس در گره‌های نهایی کلماتی با ساختارهای مشابه و با تشابه واجها وجود دارند. بنابراین منطقی است که تبدیلات واجی بصورت مشابهی رخ دهند. نتایج بدست آمده این مسأله را تأیید می‌نمایند. همانطور که در شکل ۴ دیده می‌شود. الگوی تبدیلات واجی پیش‌بینی شده اول بیان می‌نماید که واج پنجم می‌تواند حذف شود و واج سوم می‌تواند جایگزین شود این دو تغییر در ساختار واجی کلمه، با اخذ اطلاعات مناسب در مورد ماهیت واجهای کناری واج پنجم و واج سوم پیش‌بینی شده‌اند.

## ۶- قواعد تلفظی

قواعد تلفظی بصورت $L\underline{F}R \rightarrow O$ نمایش داده می‌شوند. $O, R, F, L$ بیانگر دنباله‌های واجی هستند. تفسیر چنین قاعده‌ای به این صورت است که اگر در یک دنباله واجی مرجع کلمه دنباله واجی $F$ وجود داشته باشد و اگر این دنباله توسط دنباله متنی سمت چپ $L$ و دنباله واجی سمت راست $R$ احاطه شده باشد، می‌تواند بصورت واقعی به شکل $O$ تلفظ یا بیان یا بازشناسی گردد. ترکیب $LFR$ اصطلاحاً دنباله شرط قاعده تلفظی نامیده می‌شود. چرا که وجود این دنباله واجی شرط اصلی برای اعمال قاعده است. وقتی چنین قواعدی در مرز کلمات اعمال می‌گردند. یک برچسب که نمایانگر قاعده اعمال شده است، باید به گونه تولید شده الصاق گردد تا گونه‌های تلفظی سازگار به دنبال یکدیگر بازشناسی گردند. فرآیند یادگیری قواعد و اعمال آنها به مواضع مجاز کلمات (که توسط درخت‌های تصمیم تعمیم یافته تشخیص داده شده‌اند) در این مقاله شبیه رویکرد بکار رفته در مرجع [۱۴] می‌باشد. شکل ۹ نیز مراحل استخراج قواعد تلفظی و تهیه لیستی با ارزش از قواعد را نشان می دهد.




شکل ۹- مراحل استخراج قواعد تلفظی واجی؛ این قواعد در مرحله پایانی فرایند تولید گونه های تلفظی در مدلهای ترکیبی مورد استفاده قرار می‌گیرند.

همانطور که در شکل مشخص است، بعد از همردیف سازی گونه های واقعی با مرجع الگوریتم از نواحی اختلاف دیده شده بین نسخه های همردیف سازی شده قواعد را استخراج می نماید. این قواعد بعدا با توجه به احتمال وقوع آنها هرس می شوند. تا در نهایت لیستی از قواعد تلفظی با ارزش تهیه گردد.

اگر برای تولید گونه‌های تلفظی کلمات فقط از قواعد استفاده شود، یک ضعف مهم در گونه‌های تولیدی وجود خواهد داشت. این ضعف ناشی از این واقعیت است که قواعد تلفظی دنباله واجی محدودی به عنوان شرط لازم برای اعمال به دنباله واجی کلمه دارند. این بدان معناست که این قواعد به اطلاعات در سطح کلمه مانند ساختار واجی آن، موضع تکیه در کلمه و نرخ صحبت اهمیت نمی‌دهند. حال آنکه این ویژگیها در تولید گونه‌های تلفظی کلمات بسیار با اهمیت می‌باشند. اما حسن قواعد تلفظی در مدلسازی تنوعات تلفظی امکان استخراج آنها از دادگان با حجم متوسط می‌باشد.

## ۷- نتایج آزمایشات انجام شده روی مدل تلفظی هیبرید

دو آزمایش برای اندازه‌گیری کارآیی مدل‌های ایستا و پویا ترتیب داده شده است. در اولین آزمایش، گونه‌های تولید شده برای هر کلمه توسط مدل، با گونه‌های واقعی آن کلمه که در دادگان دیده شده‌اند، همردیف‌سازی شدند. همردیف‌سازی با استفاده از الگوریتم استاندارد برنامه‌ریزی پویای مؤسسه NIST انجام می‌گیرد. سپس، فواصل بین نسخه‌های همردیف‌سازی شده، نسبت به طول دنباله‌های همردیف‌سازی شده، به‌هنجار می‌شوند. هر چه میانگین فواصل به‌هنجار شده کمتر باشد، بدین معناست که گونه‌های تولید شده توسط مدل به گونه‌های واقعی نزدیکتر هستند، و کارآیی مدل، بالاتر است. برای این آزمایش از بخشی از فارس‌دات بعنوان دادگان آزمون بهره گرفته شد، که شامل حدود ۲۰۰۰۰۰ کلمه است. در این آزمایش گونه‌های تولید شده توسط مدل‌های هیبرید ایستا و پویا با گونه‌های تولید شده توسط قواعد تلفظی و نیز با دنباله‌های واجی مرجع کلمات مقایسه شده‌اند. نتایج در جدول ۳ ارائه شده‌اند. لازم به توضیح است در مدل‌های ترکیبی پویا نسخه‌های تولیدی مدل با نسخه‌های واقعی در شرایط یکسان نرخ گفتار مقایسه شده‌اند.

جدول ۳- میانگین فواصل به‌هنجار شده بین نسخه‌های همردیف سازی شده گونه‌های تولید شده توسط مدل با گونه‌های واقعی برای تمام کلمات دادگان.

| میانگین فواصل بهنجار شده بین نسخه‌های همردیف شده گونه‌های تولیدی مدل و گونه‌های واقعی | مدل‌های استفاده شده در تولید گونه‌های تلفظی کلمات |
|---|---|
| ۰/۲۳ | مدل‌های ترکیبی پویا |
| ۰/۲۵ | مدل‌های ترکیبی ایستا |
| ۰/۲۹ | قواعد تلفظی |
| ۰/۳۴ | نسخه‌های مرجع |

در آزمایش بعدی خروجی مدلها را، که همان گونه‌های تلفظی کلمات هستند، در واژگان سیستم بازشناسی گفتار فارسی "شنوا" [۲۹]، قرار دادیم، تا آزمایشی برای اثر استفاده از گونه‌های تلفظی کلمات در فرآیند بازشناسی کلمات داشته باشیم. سیستم "شنوا" یک سیستم بازشناسی گفتار است که در مرکز تحقیقات پردازش هوشمند علائم بعنوان فاز اول پروژه بزرگ طراحی یک سیستم خودکار بازشناسی گفتار پیوسته فارسی، طراحی و ساخته شده است. در این آزمایش یک واژگان ۱۰۰۰ کلمه‌ای برای سیستم درنظر گرفته شد. بازشناسی واجها در این سیستم توسط یک ساختار ترکیبی شامل شبکه‌های عصبی و موتورهای قاعده پایه انجام می‌گیرد. بعد از جستجوی واژگانی و اعمال یک الگوریتم شبه ویتربی برای یافتن ۱۰۰ بهترین عبارت، یک بلوک امتیاز دهی مجدد که از مدلهای HMM واجهای فارسی بهره می‌گیرد، بهترین عبارت را بعنوان عبارت بازشناسی شده انتخاب می‌نماید. ویرایشی از "شنوا" که در آزمایشات از آن بعنوان سیستم پایه، استفاده شد، از هیچ مدل زبانی برای کاهش نرخ خطای بازشناسی کلمات بهره نمی‌گیرد، و نرخ خطای بازشناسی کلمات در آن ۵۳٪ است. برای تولید گونه‌های تلفظی کلمات از بخشی ۲۵ ساعته از دادگان، بعنوان دادگان آزمون بهره گرفتیم. دادگان آموزش و آزمون کاملا جدا می‌باشند. با حذف گونه‌های تلفظی کم احتمال کلمات واژگان، بطور میانگین ۲/۵ گونه تلفظی برای هر کلمه در واژگان قرار داده شد. در این آزمایش کارآیی واژگان حاوی گونه‌های تولیدی توسط مدل‌های هیبرید درخت تصمیم/قاعده تلفظی در دو نوع ایستا و پویا با کارآیی واژگان سیستم پایه شامل گونه های مرجع و نیز کارآیی واژگان حاوی گونه‌های تولیدشده توسط فقط اِعمال قواعد تلفظی (بدون استفاده از درخت تصمیم تعمیم یافته) مقایسه شده اند. بهبودها بصورت کاهش نسبی نرخ خطای بازشناسی کلمات نسبت به ۵۳٪ نرخ خطای سیستم پایه و ۵۰/۲٪ نرخ خطای مربوط به سیستم حاوی گونه های تولیدی با قواعد محاسبه و در جدول ۴ آمده‌اند. اگرچه نتایج گزارش شده تا حد زیادی وابسته به کارآیی سیستم بازشناسی گفتار پیوسته پایه، زبان مورد آزمایش، حجم واژگان و میانگین تعداد گونه‌های برای هر کلمه هستند، اما مقایسه این نتایج با نتایج گزارش شده در تحقیقات مشابه، کارآیی روش ما را تایید می‌کند.

جدول ۴- کاهش نسبی درصد نرخ خطای بازشناسی کلمات، با استفاده از گونه‌های تلفظی کلمات در واژگان در مقایسه با استفاده از فقط نسخه‌های واجی مرجع

| کاهش نسبی نرخ خطای بازشناسی کلمات در مقایسه با واژگان حاوی گونه های تولید شده با قواعد تلفظی | کاهش نسبی نرخ خطای بازشناسی کلمات در مقایسه با واژگان حاوی نسخه‌های واجی مرجع کلمات | مدلهای استفاده شده در تولید گونه‌های تلفظی کلمات برای استفاده در واژگان. |
|---|---|---|
| ٪۶/۱ | ٪۱۰/۳ | مدل هیبرید آماری پویا |
| ٪۳/۳ | ٪۸/۱ | مدل هیبرید آماری ایستا |

مقایسه کمی روشهای مختلف مدلسازی تلفظ بدلیل اینکه ارزیابی و آزمایش این روشها روی سیستم های بازشناسی گفتار متفاوت و با حجم واژگان مختلف انجام میشوند و نیز مدل ها با استفاده از پایگاههای داده مختلف در زبان های مختلف آموزش یافته اند دشوار است و اساسا مقایسه مستقیم روشهای مختلف در این حوزه کار صحیحی نمیباشد. با اینحال بنظر می رسد مروری بر نتایج برخی تحقیقات معتبر دیگر در این حوزه سودمند است و می تواند ارزش و جایگاه تحقیق گزارش شده در این مقاله را بیشتر روشن سازد.

فوکودا و همکارانش در [۲۳] از شبکه عصبی برای تولید گونه های تلفظی کلمات استفاده نمودند. در این کار دیالوگ هایی از ۲۳۰ گوینده شامل ۱۰۰ مرد و ۱۳۰ زن برای آموزش مدل های آکوستیکی و تلفظی استفاده گردید و تلفظ های مرجع و بازشناسی شده در واحد های ۵ واجی با تعداد کل ۱۲۰۰۰۰ نمونه بعنوان جفت ورودی-خروجی های آموزش شبکه عصبی استفاده شدند. آنها به ۳/۴٪ بهبود در درصد نرخ خطای بازشناسی کلمات دست یافتند، در حالیکه درصد نرخ خطای بازشناسی کلمات در سیستم پایه مورد استفاده ایشان ۳۴/۵٪ گزارش شده بود.

کرملی و مارتنز در [۱۳] با استفاده از قواعد تلفظی و ایجاد اولویت بندی در مجموعه قواعد و نیز با تعریف و بکارگیری قواعد بازدارنده به تولید گونه های تلفظی کلمات پرداختند. ایشان برای استخراج قواعد تلفظی از پایگاه داده TIMIT شامل ۶۳۰۰ جمله استفاده نمودند و با بکارگیری گونه های تولیدی در یک سیستم بازشناس گفتار توانستند نرخ خطای بازشناسی کلمات را به میزان ۱/۱۴٪ بهبود دهند؛ در حالیکه خطا در سیستم پایه ۸/۳۹٪ گزارش شده بود.

فوسلر در [۲۲] با استفاده از قواعد تلفظی به مدلسازی تنوعات تلفظی واجی در سطح هجا و کلمه پرداخت. مدلهای فوسلر گونه های بکار گرفته شده در سیستم بازشناس گفتار را بصورت پویا با توجه به کلمات یا واجهای اطراف، نرخ صحبت و قابلیت پیش بینی کلمه تغییر می دهند. فوسلر روی سیستم بازشناس گفتار ICSI به بهبود نسبی نرخ خطای ۵-۴٪ دست یافت.

ما در بهترین وضعیت به کاهش ۵/۹٪ در نرخ خطای بازشناسی کلمات دست یافتیم. با توجه به اینکه نرخ خطای سیستم پایه ۵۳٪ بوده است این نتیجه معادل با کاهش نسبی نرخ خطای بازشناسی معادل با ۱۰/۳٪ می باشد که نتیجه قابل توجهی در مقایسه با کارهای دیگر است. این در حالیست که ما برای آموزش مدلهای خود از یک پایگاه داده با حجم متوسط شامل ۲۵ ساعت داده گفتاری استفاده کرده ایم.

## ۸- نتیجه‌گیری

در این مقاله، مدل ترکیبی درخت تصمیم/ قاعده جدیدی را برای تولید گونه‌های تلفظی کلمات معرفی نمودیم. درخت تصمیم تعمیم یافته، به نحو مناسبی می‌تواند خصوصیات مشابه واج‌های قرار گرفته در گروه‌های واجی یکسان را مدل نماید. نتایج آزمایشات نشان می‌دهند که گونه‌های تلفظی که توسط مدل‌های ترکیبی ساخته شده‌اند، نسبت به گونه‌های دیگر به گونه‌های تلفظی واقعی دیده شده در دادگان نزدیکتر می‌باشند. همچنین استفاده از گونه‌های تولید شده توسط مدل ترکیبی، در واژگان یک سیستم بازشناسی گفتار، باعث بهبود کارآیی آن شده و این گونه کارآئی مدل در عمل نیز دیده می‌شود. هر دو آزمایش کارآیی بیشتر مدل‌های ترکیبی را نسبت به استفاده از تنها قواعد تلفظی تایید می‌نمایند. این کارآیی بیشتر، نتیجه ظرفیت مدل‌های ترکیبی در مدلسازی ساختار کلی کلمه می‌باشد. به این ترتیب نتایج این تحقیق نشان می‌دهند که استفاده از درخت تصمیم تعمیم یافته، راه حلی برای مشکل کمبود حجم دادگان گفتاری است، زمانی که می‌خواهیم ساختار کلی کلمه را هم در موقع مدلسازی تنوعات تلفظی کلمات مورد توجه قرار دهیم.

## مراجع


[1] H. Strik, and C. Cucchiarini, "Modeling Pronunciation Variation for ASR: A survey of the literature," *Speech Communication,* Vol. 29, pp: 225-246, 1999.

[2] M. Eskenazi, "Trends in Speaking Styles Research", *In Proceedings of the Eurospeech, Berlin,* pp. 501-509, 1993.

[3] J. Laver, *Principles of Phonetics,* Cambridge University Press, Cambridge, 1994.

[4] A. Bell, "Language Styles as Audience Design", *Language in Society,* Vol. 13, No. 2, pp: 145:204, 1984.

[5] H. Giles, *Speech Style and Social Evaluation,* Cambridge University Press, Cambridge, 1975.

[6] D. Jurafsky, A. Bell, M. Gregory, and W. D. Raymond, "The Effects of Language Model Probability on Pronunciation Reduction", *In Proceedings of the ICASSP, Utah,* Vol. 2, pp. 801-804, 2001.

[7] J. Barnett, "A Phonological Rule Compiler", *In Proceedings of the IEEE Symposium on Speech Recognition, Carnegie Mellon University, Pittsburgh,* pp: 188-192, 1974.

[8] E. Fosler, I. Amdal, and H. J. Kuo, "On the Road to Improved Lexical Confusability Metrics", *In Proceedings of the Pronunciation Modeling and Lexicon Adaptation, Colorado,* 2002.

[9] L. Bahl, J. Baker, P. Cohen, F. Jelinek, B. Lewis, and R. Mercer, "Recognition of a Continuously Read Natural Corpus", *In Proceedings of the ICASSP-78*, pp. 422-424, 1978.

[10] C. Wooters, and A. Stolcke, "Multiple Pronunciation Lexical Modeling in a Speaker Independent Speech Understanding System", *In Proceedings of the ICSLP-94*, pp. 1363-1366, 1994.

[11] T. Imai, A. Ando, and E. Miyasaka, "A New Method for Automatic Generation of Speaker-dependent Phonological Rules", *In Proceedings of the ICASSP-95*, pp. 864-867, 1995.

[۱۲] ب. وزیرنژاد، **مدلسازی تنوعات تلفظی در سیستم بازشناسی گفتار پیوسته فارسی**، پایان نامه کارشناسی ارشد مهندسی پزشکی، دانشکده مهندسی پزشکی، دانشگاه صنعتی امیرکبیر، ۱۳۸۲.

[13] N. Cremelie, and J. P. Martens, "In Search of Better Pronunciation Models for Speech Recognition", *Speech Communication,* Vol. 29, pp. 115-136, 1999.

[14] N. Cremelie and J. Martens "Automatic Rule-Based Generation of Word Pronunciation Networks", *In Proceedings of the Eurospeech*, pp. 2459-2462, 1997.

[15] Q. Yang, and J. P. Martens, "On the Importance of Exception and Cross-Word Rules for the Data-Driven Creation of Lexica for ASR", *In Proceedings of the 11th ProRisc Workshop, Veldhoven*, pp. 589-593, 2000.

[16] Q. Yang, and J. P. Martens, "Data-Driven Lexical Modeling of Pronunciation Variation for ASR", *In Proceedings of the ICSLP, Beijing*, 2000.

[17] G. Tajchman, E. Fosler, and D. Jurafsky, "Building Multiple Pronunciation Models for Novel Words Using Exploratory Computation Phonology", *In Proceedings of the 4th Eurospeech, Madrid,* pp. 2247-2250, 1995.

[18] . Wester, "Pronunciation Modeling for ASR-Knowledge-based and Data-derived Methods", *Journal of Computer Speech and Language,* vol. 17, pp. 69-85, 2003.

[19] P. A. Jande, "Pronunciation Variation Modeling using Decision Tree Induction from Multiple Linguistic Parameters", *In Proceedings of the FONETIK, Stockholm University*, pp. 12-15, 2004.

[20] H. Yu, and T.Schultz, "Enhanced Tree Clustering with Single Pronunciation Dictionary for Conversational Speech Recognition", *In Proceedings of the Eurospeech, Geneva*, pp. 1869-2596, 2003.

[21] E. Fosler-Lussier, "Contextual word and syllable pronunciation models," *In Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, Keystone, Colorado,* 1999.

[22] E. Fosler-Lussier, *Dynamic Pronunciation Models for Automatic Speech Recognition,* Ph.D. thesis, University of California, Berkeley, 1999.

[23] T. Fukada, T. Yoshimura, Y. Sagisaki, "Automatic Generation of a Multiple Pronunciations Based on Neural Networks", *Speech Communication,* Vol. 27, Issue 1, pp. 63-73, 1999.

[24] K. Chen, and M. Hasegawa-Johnson, "Modeling Pronunciation Variation Using Artificial Neural Networks for English Spontaneous Speech", *In Proceedings of the ICSLP*, 2004.

[25] T. Hazen, L. Hetherington, L. Shu, K. Livescu, "Pronunciation Modeling Using a Finite State Transducer Representation", *Speech Communication,* Vol. 46, Issue 2, pp. 189-203, 2005.

[26] M. Davel, E. Barnard, "Bootstrapping Pronunciation Dictionaries", *South African Journal of Science,* Vol. 102, No. 7-8, pp. 322-328, 2006.

[27] N. Mirghafori, E. Fosler, and N. Morgan, "Why is ASR Harder for Fast Speech and What Can We Do about it?", *In Proceedings of the IEEE Snowboard Workshop, Utah,* Dec. 1995.

[28] M. Bijankhan, and M. J. Sheikhzadegan, "FARSDAT-the Farsi Spoken Language Database", *In Proceedings of the Fifth International Conference on Speech Sciences and Technology, Perth*, Vol. 2, pp.826-829, 1994.

[29] F. Almasganj, et al., "SHENAVA-1; A Persian Spontaneous speech recognizer", *In Proceedings of the Tenth International Conference on Electrical Engineering, Tehran*, pp.101-106, 2001.

[۳۰] ع. م. حقشناس، **آواشناسی**، چاپ چهارم، انتشارات آگاه، تابستان ۱۳۷۴.

[۳۱] ه. میلانیان، ر. سید حسینی، **زبانشناسی مجموعه مقالات**، وزارت فرهنگ و ارشاد اسلامی، ۱۳۸۱.

**بهرام وزیرنژاد** در زمستان ۱۳۵۷ در تهران متولد شد. او در سال ۱۳۷۹ درجه کارشناسی خود را در رشته مهندسی پزشکی از دانشگاه علوم پزشکی شهیدبهشتی تهران اخذ نمود. متعاقب آن در سال ۱۳۸۲ موفق به اخذ درجه کارشناسی ارشد در رشته مهندسی پزشکی گرایش بیوالکتریک از دانشگاه صنعتی امیرکبیر گردید. او از سال ۱۳۸۲ دانشجوی مقطع دکتری تخصصی در رشته مهندسی پزشکی گرایش بیوالکتریک در دانشگاه صنعتی امیرکبیر می‌باشد. همچنین از سال ۱۳۸۱ تا کنون پژوهشگر مرکز تحقیقات پردازش هوشمند علائم در پروژه های مختلف پردازش گفتار بوده است. او دارای سابقه تدریس ریاضیات و الکترونیک در دانشکده مهندسی پزشکی، دانشگاه آزاد اسلامی واحد علوم و تحقیقات است. از او تا کنون ۴ مقاله در نشریات علمی پژوهشی و ۶ مقاله در کنفرانسهای معتبر داخلی و خارجی به چاپ رسیده است. زمینه کارهای تحقیقاتی او تا کنون پردازش سیگنال گفتار، بازشناسی گفتار پیوسته و روش‌های بازشناخت الگو بوده است. آدرس پست الکترونیکی ایشان: bvazirnezhad@aut.ac.ir

**فرشاد الماس‌گنج** در سال ۱۳۶۳ در رشته برق گرایش الکترونیک در مقطع کارشناسی از دانشگاه صنعتی امیرکبیر فارغ‌التحصیل گردیده است. سپس در همین گرایش دوره کارشناسی ارشد خود را در سال ۱۳۶۷ به پایان رسانید، و به سمت عضو هیئت علمی دانشگاه صنعتی امیرکبیر مشغول بکار گردید. بعد از تاخیری کوتاه، تحصیل خود را در رشته برق گرایش مهندسی پزشکی در دانشگاه تربیت مدرس ادامه داد؛ و در سال ۱۳۷۷ به درجه دکتری دست یافت. از آن زمان تا کنون در دانشکده مهندسی پزشکی دانشگاه صنعتی امیرکبیر با سمت استادیاری حضور دارد. زمینه تخصصی مورد علاقه او پردازش سیگنال و خصوصا پردازش انواع سیگنال‌های گفتاری است. آدرس پست الکترونیکی ایشان: almas@aut.ac.ir

[1] Co-articulation
[2] Lexicon
[3] Pronunciation Variants

[4] Confusion
[5] Randolph
[6] Riley
[7] Wooters
[8] Stolcke
[9] Schmid
[10] Sloboda
[11] Imai
[12] Humphries
[13] Tajchman
[14] Crème lie, Martens
[15] Alignment
[16] Fosler
[17] Wooters, Stolke
[18] Fukada, Sagisaki
[19] Chen, Hasegawa
[20] Pseudo-phonemes
[21] Generation restriction rules
[22] Hybrid
[23] Static
[24] Dynamic
[25] likelihoods