نشریه پژوهش‌های زبان‌شناسی تطبیقی
سال پنجم- شماره ۱۰- پاییز و زمستان ۱۳۹۴

مدل‌سازیِ نوای گفتارِ کانونی در فارسی: رویکردی تولیدی ـ نقش‌گرا*

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

چکیده

مقالهٔ حاضر تلاشی است در جهت مدل‌سازیِ نوای گفتارِ کانونیِ فارسی که با اتخاذ رویکردی تولیدی ـ نقش‌گرا یعنی رمزگذاری موازی و تقریبِ هدف (PENTA) انجام شده است. داده‌های مورد استفاده برای مدل‌سازی شامل ۱۵۰ پاره‌گفتار است که در شرایط مختلف کانونی و غیرکانونی تولید شده‌اند. در راستای رسیدن به این هدف، از بازسازی‌کنندهٔ PENTAtrainer2 تحتِ نرم‌افزار پِرَت (Praat) استفاده شده است. این بازسازی‌کننده در قالبِ رویکرد PENTA، اهدافِ زیروبمیِ مقوله‌ای را بهینه می‌کند که هر یک از این اهداف با نقش‌های ارتباطی خاصی مرتبط می‌باشند. ارزیابی عینی از مقایسهٔ منحنی بسامدپایهٔ بازسازی‌شده با منحنی بسامدپایهٔ طبیعی حاکی از آن است که منحنی بازسازی‌شده با خطای جذرِ میانگینِ مربعات ۱/۹۴ و میزان ضریب همبستگی ۰/۸۴، به منحنی طبیعی شباهت بسیار دارد. همچنین ارزیابی ذهنی از جایگاه کانون و همچنین قضاوت آزمودنی‌ها از میزان طبیعی‌بودن صداهای بازسازی‌شده، نشان‌دهندهٔ شباهت بسیار زیاد هر دو منحنیِ طبیعی و بازسازی‌شده است.

کلیدواژه‌ها: رمزگذاری موازی، تقریب هدف، بازسازی، کانون، بسامدپایه.

* این مقاله برگرفته از رسالهٔ دکترای نگارنده است.

۱- دانشجوی دکتری پژوهشگاه علوم انسانی و مطالعات فرهنگی ✉| m.taheri@ihcs.ac.ir
۲- استاد پژوهشگاه علوم انسانی و مطالعات فرهنگی ✉| s_m_assi@ihcs.ac.ir
۳- دانشیار دانشگاه صنعتی شریف ✉| sameti@sharif.edu
۴- استاد دانشگاه تهران ✉| mbjkhan@ut.ac.ir

تاریخ دریافت: ۹۴/۰۳/۲ تاریخ پذیرش: ۹۴/۰۹/۳۰

مقدمه

نوای گفتار به عنوان بخش اصلی و جدایی‌ناپذیر گفتار همواره چالشی برای علم فناوری گفتار بوده است. برای نمونه، تولید نوای گفتار طبیعی در سیستم‌های تبدیل متن به گفتار[1] هنوز به عنوان یک مسئلهٔ حل نشده باقیمانده است. پاسخی قانع‌کننده به این مسئله، نه تنها کمک شایانی به پیشرفت فناوری گفتار می‌کند بلکه درک فرایند گفتار را برای دیگر محققان در این زمینه هموار می‌کند. در سالیان گذشته، در راستای حل این مسئله زبانشناسان و متخصصین پردازش گفتار نظریه‌ها و الگوهای مختلفی ارائه کرده‌اند. در اوایل دههٔ هشتاد از قرن بیستم، رویکرد خودواحد –وزنی[2] به‌عنوان معتبرترین رویکردِ زبان‌شناختی به عناصر زبرزنجیری، به معرفی چارچوبی پرداخت که با استفاده از آن بتوان الگویی نظری برای آهنگ گفتارِ زبان‌های مختلف ارائه داد. نخستین بار پیرهامبرت[3] (۱۹۸۰) چارچوب کلی و مفاهیم نظری این رویکرد را معرفی کرد. این رویکرد بعدها دچار تعدیل‌هایی شد اما مبانی اولیه و اصول نظری آن بدون تغییر تا به امروز باقی‌مانده است. در این چارچوب از دو نواخت بالا[4] و پایین[5] برای توصیف آهنگ استفاده شده است که در قالب نواخت‌های مرزنما[6] و تکیه‌های زیروبمی[7] تجلی پیدا می‌کنند. حاصل تحقیقات گسترده در این زمینه معرفی نظامی برچسب‌گذاری است که نخستین بار برای زبان انگلیسی امریکایی تحت عنوان برچسب‌گذاری نواخت‌ها و فاصله‌نماها[8] تدوین شده است (سیلورمن[9] و همکاران، ۱۹۹۲). یکی از دلایل معرفی چنین نظامی به منظور بهره‌برداری از آن در مدل‌سازی نوای گفتار در سیستم‌های TTS بوده است که در دههٔ پایانیِ قرن پیش، توجه بسیاری از محققینِ این حوزه را به خود جلب کرد و تلاش‌های بسیار در این زمینه به‌ویژه برای زبان انگلیسی صورت گرفت. در فارسی، طاهری اردلی و خرم (۱۳۹۱) در پژوهشی با استفاده از نظامِ توصیفیِ ارائه‌شده در قالبِ رویکردِ خودواحد ـ وزنی که سادات‌تهرانی (۲۰۰۷) پیشنهاد کرده است، به پیاده‌سازی نوای گفتار در یک سیستم بازسازیِ گفتار[10] با پشتوانهٔ دادگانی مشتمل بر ۱۳۰۰ پاره‌گفتار پرداختند. این سیستم که پیاده‌سازی آن مبتنی بر مدل مخفی مارکوف[11] بوده است، کیفیتِ خروجی آن با میانگینِ امتیازاتِ نظردهی[12] ۴.۶ ارزیابی شده است.

1. text-to-speech system (TTS)
2. Auto segmental-Metrical
3. Pierrehumbert
4. high tone (H)
5. low (L)
6. boundary
7. pitch accent
8. Tones and Break Indices (ToBI)
9. Silverman
10. speech synthesis
11. Hidden Markov Model (HMM)
12. Mean Opinion Score (MOS)

رویکرد فوجی‌ساکی[1] چارچوبی دیگر است که با نگرشی نسبتاً متفاوت، به صورت جمع‌آثار[2] به توصیف دقیقی از منحنی بسامدپایه[3] پرداخته است (فوجی‌ساکی و ناگاشیما[4]، ۱۹۶۹؛ فوجی‌ساکی و هیروسه[5]، ۱۹۸۴). ورودی این الگو در قالب تکانه‌ها[6] و توابع پله‌ای[7] است که به ترتیب برای تولید گروه[8] و تکیهٔ زیروبمی به عنوان دو مولفهٔ اصلی این رویکرد، مورد استفاده قرار می‌گیرند. هر گروه با یک تکانه آغاز می‌شود که با عبور از صافی درجهٔ دوم میرا[9] باعث می‌شود تا منحنی بسامدپایه به ارزش حداکثری خود برسد و سپس رو به زوال حرکت کند. اما تکیهٔ زیروبمی با استفاده از یک تابع پله‌ای آغاز می‌شود و زمانی که این تابع از صافی عبور می‌کند پاسخ‌هایی را تولید می‌کند. از جمع‌آثار و ترکیب همزمان دو مولفهٔ گروه و تکیهٔ زیروبمی به همراه ارزش مبنایِ[10] بسامدپایه، منحنیِ نهایی بدست می‌آید. این رویکرد برای زبان‌های بسیاری مورد آزمایش قرار گرفته است. در فارسی، نم‌نبات و کوچاری (۱۳۸۶) با استفاده از پایگاه‌دادگانِ[11] فارس‌دات بزرگ (بی‌جن‌خان، ۱۹۹۴) به استخراج خودکار پارامترهای فوجی‌ساکی پرداختند. نتایج تحقیق آنها نشان می‌دهد که بازسازی نوای گفتار فارسی با حداقل پارامترها یعنی بدون نیاز به فرمانِ[12] تکیهٔ زیروبمی نیز عملی است. در پژوهش آنها، مقایسهٔ منحنی طبیعی و منحنی تولیدی، ضریب همبستگیِ[13] حدود ۹۷ درصد و خطایِ جذرِ میانگین مربعاتِ[14] برابر با ۴/۶۵ هرتز را نشان می‌دهد.

یکی دیگر از رویکردها با نگرشی مهندسی به نوای گفتار، تیلت[15] است (تیلور، ۱۹۹۲؛ ۲۰۰۰). این الگو، آهنگ گفتار را به عنوان مجموعه‌ای از رخدادهای نواختی در نظر می‌گیرد. در این الگو، به مانند رویکرد خودواحد ـ وزنی از دو نوع رخداد یعنی تکیهٔ زیروبمی و نواخت مرزنما استفاده می‌شود اما برخلاف خودواحد ـ وزنی که از مجموعه‌ای ثابت از مقوله‌ها بهره می‌گیرد، از پارامترهای پیوسته استفاده می‌کند. هر یک از این رخدادها از دو بخش خیز[16] و افت[17] تشکیل شده است و بین آنها خطوط مستقیمی وجود دارد که خطوط اتصال[18] نامیده می‌شوند. تنوع و تفاوتی که در تکیهٔ زیروبمی و نواخت مرزنما وجود دارد تحت تاثیر اندازهٔ

1. Fujisaki
2. Super positional
3. F0 contour
4. Nagashima
5. Hirose
6. impulse
7. step function
8. phrase
9. damped second order filter
10. baseline
11. database
12. command
13. correlation
14. Root Mean Square Error (RMSE)
15. Tilt
16. rise
17. fall
18. connection

نسبیِ خیز و افت‌هاست و اینکه چگونه این رخدادها با عناصر زنجیری گفتار برهم‌نهاده[1] می‌شوند. در مجموع، Tilt شش پارامتر را شامل می‌شود. چهار پارامتر که شکل رخدادها را توصیف می‌کند و دو پارامتر دیگر، چگونگی برهم‌نهادگی رخدادها با عناصر زنجیری را نشان می‌دهد. در راستای بکارگیری این رویکرد، نمنبات و همکاران (۱۳۸۴) در پژوهشی به توصیف F_0 در فارسی پرداختند سپس با استفاده از درخت طبقه‌بندی و رگرسیون[2] به مدل‌سازیِ پارامترها همت گماشتند که در نهایت مقایسهٔ منحنی تخمینی و طبیعی برای مجموعهٔ آزمون[3] ، ضریب همبستگیِ ۶۱/۱ درصد و خطایِ جذرِ میانگین مربعاتِ برابر ۲۵/۳۸۶ هرتز را نشان می‌دهد.

از رویکردهایِ متاخر در این زمینه، الگوی رمزگذاری موازی و تقریب هدف[4] است که پیاده‌سازی آن را می‌توان در نگاهی کمی، یعنی تقریبِ کمیِ هدف[5] مشاهده کرد (ژو[6]، ۲۰۰۵؛ پروم ـ آن[7] و همکاران، ۲۰۰۹؛ ژو و پروم ـ آن، ۲۰۱۴). پیش از این، قدرت پیش‌بینی بالای بازسازیِ منحنی F_0 با استفاده از این الگو در زبان‌های انگلیسی، ماندرین، ژاپنی و تایلندی گزارش شده است (ژو و پروم ـ آن، ۲۰۱۴) در مقالهٔ حاضر، با اتخاذ الگوی مذکور و با استفاده از مجموعه دادگانی بالغ بر ۱۵۰ پاره‌گفتار در شرایط مختلف کانونی، تحلیل و بازسازیِ نوای گفتار کانونیِ جملات خبری در فارسی، مورد مطالعه و بررسی قرار می‌گیرد. لازم به ذکر است که برای نخستین بار است که مدل‌سازیِ نوای گفتار با استفاده از این رویکرد در فارسی انجام می‌گیرد.

اهمیت انجام این پژوهش برای نگارندگان از آن جهت است که در صورت موفقیت در دستیابی به کیفیتِ مطلوبی از منحنی بسامدپایهٔ تولیدی، برخلاف رویکردهای دیگر، می‌توان به‌راحتی از آن در سیستم‌های TTS بهره گرفت.

در ادامه، در بخش ۲ چارچوب کلی PENTA توضیح داده خواهد شد که خود مشتمل بر سه زیربخشِ نقش‌های ارتباطی[8]، تقریب هدف و PENTAtrainers است. در بخش ۳ به طور مختصر به کانون نوایی خواهیم پرداخت. در بخش ۴ به روش‌شناسی تحقیق اشاره شده است که دربردارندهٔ توضیحاتی پیرامون داده‌های تحقیق، برچسب‌گذاری نقشی، آزمون ادراکی و نتایج تحقیق است. در نهایت، بخش ۵ پیش از منابع، به بحث و نتیجه‌گیری اختصاص یافته است.

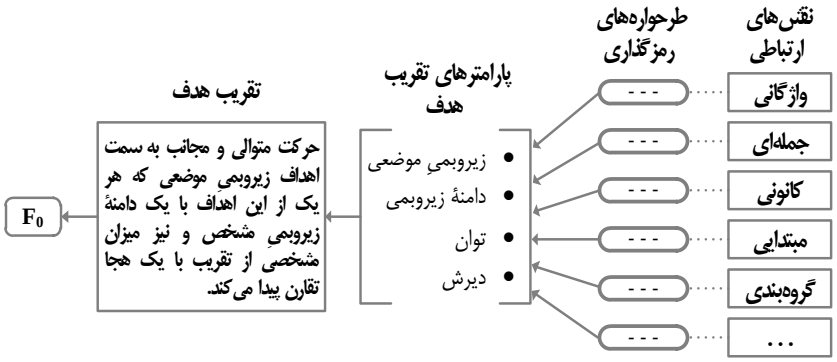
1. align
2. Classification and Regression Tree (CART)
3. test set
4. Parallel Encoding and Target Approximation (PENTA)
5. quantitative Target Approximation (qTA)
6. Xu
7. Prom-on
8. communicative function

الگوی PENTA

الگوی PENTA، چارچوبِ مورد استفاده در این تحقیق، با اتکا به نگرشی تولیدی ـ نقش‌گرا[1] به گفتار، بین معنایِ ارتباطی و نوای گفتار پیوند برقرار می‌کند (ژو، ۲۰۰۵؛ ژو و پروم ـ آن، ۲۰۱۴) این الگو از ابتدا، برروی دو جنبه از نوای گفتار یعنی نقش(کارکرد)های ارتباطی و سازوکارهای تولیدی تاکید داشته است (ژو، ۲۰۰۵). این دو جنبه هر یک به طور مجزا در ادامه توضیح داده شده است.

نقش‌های ارتباطی

در الگوی PENTA، نقش ارتباطی یک معنای ارتباطی خاص است که گوینده با بهره‌گیری از عناصر زبرزنجیری قصد انتقال آن را به شنونده دارد. همانطور که در شکل ۱ ترسیم شده است مستطیل‌های سمت راست، نقش‌های ارتباطیِ جداگانه‌ای هستند که به عنوان نیروی محرکِ الگو ایفای نقش می‌کنند. هر یک از این نقش‌ها، طرحوارهٔ رمزگذاریِ[2] منحصربفردی دارند که متشکل از پارامترهایِ اهداف زیروبمی[3] هستند. اهداف زیروبمی، کوچکترین واحدهایی هستند که با واحدهای زیروبمیِ زبان‌شناختی مانند نواخت و تکیهٔ زیروبمی در پیوندند (ژو و ونگ، ۲۰۰۱). پارامترها در مستطیلِ «پارامترهایِ تقریبِ هدف» در میانهٔ شکل ۱ آورده شده است. اهداف زیروبمی به لحاظ تولیدی از طریق تقریبِ هدف که منحنیِ F_0 نهایی را تولید می‌کند، پیاده‌سازی می‌شوند. در نتیجه، چارچوب نظریِ PENTA، تولید نوای گفتار به‌مثابه فرایندِ رمزگذاریِ *نقش‌های ارتباطی* از طریق *تقریب هدف* را توصیف می‌کند.

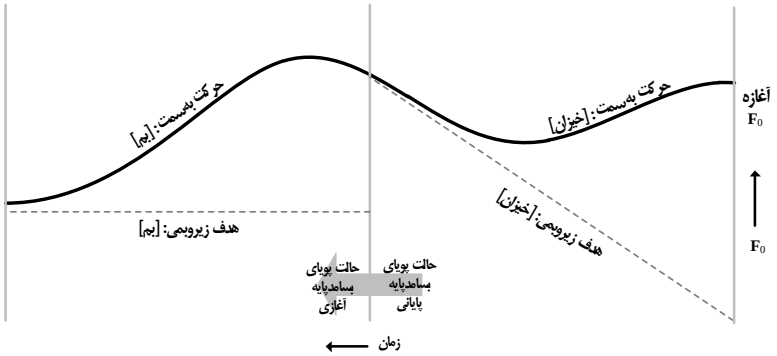


شکل ۱. طرحواره‌ای از بخش‌های مختلف چارچوب نظریِ رمزگذاری موازی و تقریبِ هدف (PENTA). این شکل برگرفته از ژو (۲۰۰۵) است.

1. articulatory - functional
2. encoding schema
3. pitch target

تقریب هدف

در شکل ۲ فرآیند تقریب هدف به تصویر کشیده است که توسط ژو و ونگ (۲۰۰۱) مطرح شده است. خط خمیدهٔ سیاه‌رنگ، منحنی F_0 را نشان می‌دهد که به صورت مجانب[1] به سمت دو هدفِ زیروبمیِ متوالی حرکت می‌کند. این دو هدف یکی پویا[2] و دیگری ایستاست[3] که در شکل زیر به صورت خطوط بریده نمایش داده شده است. در توضیحِ هدف زیروبمیِ پویا و ایستا باید اشاره کرد که نوع اول دربردارندهٔ نواخت‌های خیزان و افتان است و نوع دوم، یعنی ایستا، شامل نواختِ بالا، پایین و یا متوسط[4] است (ژو و ونگ[5]، ۲۰۰۱).



شکل ۲. فرایند تقریب هدف

منحنی سیاه نشان‌دهندهٔ منحنی F_0 است که به صورت مجانب از سمت راست به طرف دو هدف زیروبمیِ متوالی که با خط بریده نشان داده شده است حرکت می‌کند. خط خاکستری رنگِ عمودی در وسط، بیانگر مرز هجاست که از مسیر آن مرحلهٔ پویایِ F_0 پایانی، به هجای بعدی منتقل می‌شود. پیکانِ خاکستری، مسیر انتقال مرحلهٔ پویای F_0 را نشان می‌دهد. این شکل برگرفته از ژو و پروم ـ آن (۲۰۱۴) است.

این الگویِ مفهومی، اخیراً به عنوان الگوی تقریبِ کمیِ هدف توسط پروم ـ آن و همکاران (۲۰۰۹) پیاده‌سازی شده است. در تقریبِ کمیِ هدف یا همان qTA بسامدپایهٔ هر هجا با استفاده از معادلهٔ خطیِ درجهٔ دوم زیر بدست می‌آید:

$$f_0(t) = (mt + b)(c_1 + c_2 t + c_3 t^2)e^{-\lambda t} \qquad (۱)$$

1. asymptotically
2. dynamic
3. static
4. mid
5. Wang

که در آن m و b به ترتیب شیب[1] و ارتفاع[2] اهداف زیروبمی هستند و λ توان یا قدرت[3] تقریبِ هدف است. همچنین سه ضریب گذرا[4] در معادلهٔ بالا یعنی c_1، c_2 و c_3 از طریق فرمول زیر محاسبه می‌شوند:

$$c_1 = f_0(0) - b \quad (۲)$$

$$c_2 = f_0'(0) + c_1\lambda - m \quad (۳)$$

$$c_3 = (f_0''(0) + 2c_2\lambda - c_1\lambda^2)/2 \quad (۴)$$

qTA از سه پارامتر فوق یعنی m، b و λ استفاده می‌کند تا برروی هر یک از منحنی‌های F_0 در هر هجا نظارت داشته باشد. ذکر این نکته ضروری است که در الگوی مذکور، هجا به عنوان واحد پایهٔ حاملِ نوای گفتار در نظر گرفته شده است.

ابزار بازسازی‌کنندهٔ PENTAtrainers

پارامترهای هدف در qTA به شیوه‌های مختلفی بدست می‌آیند. آن‌ها را می‌توان به صورت دستی مشخص کرد و یا به صورت خودکار از طریق آموزش[5] از روی داده‌های واقعیِ گفتار بدست آورد. برای دستیابی به پارامترهای خودکار، دو برنامهٔ PENTAtrainer1 (پروم ـ آن و همکاران، ۲۰۰۹) و PENTAtrainer2 (ژو و پروم ـ آن، ۲۰۱۴) که تحت نرم‌افزار پِرَت (بورسما و وینینک[6]، ۲۰۰۱) عمل می‌کنند توسعه یافت. این دو برنامه، پارامترهای هدف را از طریق تحلیل به شیوهٔ بازسازی[7] استخراج می‌کنند. تفاوت نسخهٔ نخست و نسخهٔ دوم در نحوهٔ بهینه‌سازی[8] است. نحوهٔ عملکرد برنامهٔ PENTAtrainer1 به این نحو است که یک جستجوی جامع[9] انجام می‌دهد و همهٔ هدف‌های ممکن را در یک دامنه[10] مورد آزمایش قرار می‌دهد سپس از بین این اهداف، هدفی را انتخاب می‌کند که مناسبترین گزینه (به لحاظ شباهت به F_0 طبیعی) برای هر یک از هجاهاست. با این روش، اهداف مقوله‌ایِ با نقش‌هایِ ارتباطیِ یکسان، تنها می‌توانند از طریق میانگین‌گیریِ اهدافی که به مقوله‌های مشابه تعلق دارند بدست آیند (پروم ـ آن، ۲۰۰۹). در مقابل، PENTAtrainer2 اهدافِ مقوله‌ایِ بهینه را مستقیماً با یک جستجوی تصادفیِ[11] همه‌جانبه[12] در کل

1. slope
2. height
3. strength
4. transient coefficient
5. train
6. Boersma & Weenink
7. analysis-by-synthesis
8. optimization
9. exhaustive search
10. range
11. stochastic
12. global

پایگاه‌دادگان گفتاری بدست می‌آورد. در تحقیق حاضر با بکارگیری برنامهٔ PENTAtrainer2 اهداف مقوله‌ای به صورت خودکار از یک مجموعهٔ دادگان گفتاری استخراج شده است. در بخش بعدی به کانون نوایی می‌پردازیم که هدف نهایی از پژوهش حاضر بازسازی این جنبه از نوای گفتار فارسی است.

کانون نوایی

کانون، کارکردی ارتباطی است که به منظور برجسته‌کردن یا مورد تاکید قراردادنِ بخشی از پاره‌گفتار اعمال می‌شود و می‌تواند با استفاده از ابزارهای صرفی- نحوی از جمله اسنادسازی[۱] نمود پیدا کند. این تأکید[۲] می‌تواند با استفاده از ابزارهای نوایی نیز تحقق پیدا کند. عنصر کانونی غالباً با F_0 بیشتری نسبت به همتای غیرکانونی خود نمود پیدا می‌کنند اما شواهد متقنی نیز وجود دارد دال بر اینکه کانون در بسیاری از زبان‌های دنیا نه تنها با افزایش F_0 برروی عناصر کانونی، بلکه با تراکم دامنه زیروبمیِ عناصر پساکانونی نیز نمود پیدا می‌کند که از این پدیده تحت عنوان تراکم پساکانونی[۳] یاد می‌شود (چن[۴] و همکاران، ۲۰۰۹). طاهری اردلی و ژو (۲۰۱۲) در پژوهشی نشان دادند که در فارسی، F_0 به عنوان همبستهٔ اصلی کانون نوایی، در ناحیهٔ پیش‌کانون تغییر معناداری را نشان نمی‌دهد اما در ناحیهٔ کانون و پس‌کانون این تغییرات کاملاً معنادار است؛ یعنی، عناصر کانونی به طور ملموسی با F_0 بالاتری نسبت به همتای غیرکانونی خود نمود پیدا می‌کنند در عناصر پس‌کانونی، F_0 جملات کانونی نسبت به معادل فاقد کانون خود پایین‌تر است (شکل ۳ و ۵ نمونه‌ای از منحنیِ F_0 کانونی را در فارسی نشان می‌دهد).

در یک نگاه کلی، از چشم‌انداز رده‌شناسی نوایی[۵] ، نوای گفتار فارسی به عنوان یک زبان تکیه‌بر[۶] در سطح کلمه (فرگوسن[۷]، ۱۹۵۷؛ لازار[۸]، ۱۹۵۷؛ کهنمویی‌پور[۹]، ۲۰۰۳؛ جان[۱۰]، ۲۰۰۵) و در سطح جمله، شامل گروه‌های تکیه‌ای[۱۱] است که حاوی تکیهٔ زیروبمیِ منفرد هستند و در انتها، بسته به نوع عبارت با نواخت‌های مرزنمایِ بالا یا پایین خاتمه پیدا می‌کنند (ماهجانی، ۲۰۰۳؛ سادات‌تهرانی، ۲۰۰۷؛ ابوالحسنی‌زاده و همکاران، ۲۰۱۲).

1. clefting
2. emphasis
3. post-focus compression (PFC)
4. Chen
5. prosodic typology
6. stress language
7. Ferguson
8. Lazard
9. Kahnemuyipour
10. Jun
11. accentual phrases

روش‌شناسی پژوهش

دادگان، آزمودنی‌ها و ابزارهای تحقیق

جملهٔ محرک مورد استفاده در این تحقیق «ماها بابای نیلی‌رو لندن دیدیم» است که پنج گویشور مرد با میانگین سنی ۲۵ سال، آن را در شش حالت مختلف کانونی و غیرکانونی تولید کردند. هر یک از این حالت‌ها پنج بار تکرار شد که در مجموع ۱۵۰ پاره‌گفتار بدست آمد. یعنی: ۶ حالت کانونی X ۵ گوینده X ۵ بار تکرار = ۱۵۰ پاره‌گفتار.

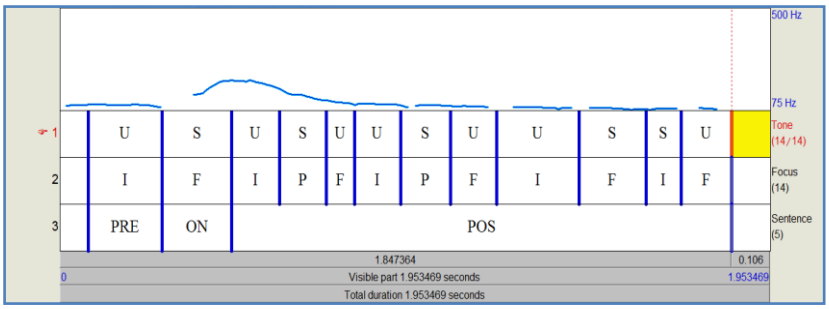
نتایج آزمون عینی[1]، یعنی ضریب همبستگی و خطای جذرِ میانگینِ مربعات، حاصل از بازسازی کلیهٔ ۱۵۰ پاره‌گفتار است. اما در آزمایش ادراکی (ذهنی[2])، به منظور مقایسهٔ نتایج جملات بازسازی‌شده با نتایج آزمایش ادراکی انجام شده با استفاده از جملاتِ طبیعی (طاهری اردلی و ژو، ۲۰۱۴) ، پاره‌گفتارهای محرکِ سه نفر از گوینده‌ها در این پژوهش استفاده شد. معیار انتخابِ گوینده‌ها بر این اساس است که این سه نفر، کمینه، بیشینه و میانهٔ انحراف معیارِ[3] بسامد پایهٔ پنج گویندهٔ حاضر در تحقیق را دارا هستند. در مجموع ۹۰ پاره‌گفتارِ بازسازی‌شده مورد استفاده قرار گرفت. یعنی: ۶ حالت کانونی X ۳ گوینده X ۵ بار تکرار = ۹۰ پاره‌گفتار. در نهایت، در این بخش یعنی آزمون ادراکی، پنج آزمودنی مرد و پنج آزمودنی زن مشارکت داشتند که به این ۹۰ پاره‌گفتار و همتای طبیعی آنها یعنی در مجموع ۱۸۰ پاره‌گفتار گوش فرا دادند و قضاوت خود را با کلیک برروی گزینه‌های مورد نظر برروی صفحهٔ رایانه ثبت کردند. این افراد هیچ‌گونه اختلال گفتاری یا شنیداری را به آزمون‌گیرنده گزارش نکردند. درضمن، هزینهٔ شرکت در آزمون به آزمودنی‌ها پرداخت شد.

به منظور انجام آزمون‌هایِ ادراکی از ExperimentMFC در نرم‌افزار پرت استفاده شد که یک شیوهٔ بسیار رایج در انجام چنین آزمون‌هایی است. در ابتدا، آزمودنی‌ها برای چگونگی انجام دو آزمون آموزش دیدند و از این اختیار برخوردار بودند تا هر تعداد از فایل‌ها را مایل هستند قبل از انجام آزمون اصلی بشنوند. سپس پاره‌گفتارهای محرک برای آزمودنی‌ها به صورت تصادفی پخش شد تا نظر خود را اعلام کنند.

برچسب‌گذاری نقشی و بازسازی نوای گفتار

با پیروی از مفروضات الگوی PENTA یعنی رمزگذاری موازی و نقش‌های ارتباطی، داده‌های تحقیق با سه لایهٔ نقشی یعنی تکیه (تکیه‌بر، بی‌تکیه)، جایگاه هجا (آغازی، ماقبل پایانی، پایانی) و شرایط کانونی (پیش‌کانون، کانون، پس‌کانون) برچسب‌گذاری شدند. شکل ۳ نمونه‌ای از این برچسب‌گذاری را در برنامهٔ پرت نشان می‌دهد.

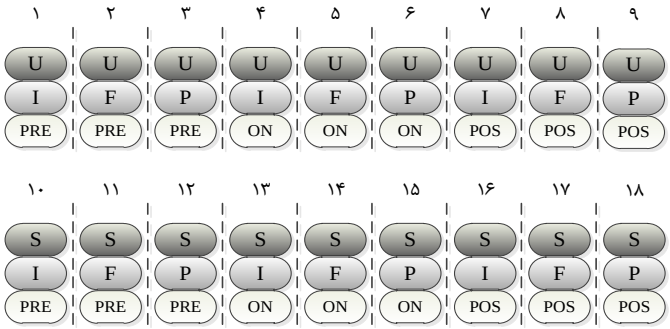
1. objective
2. subjective
3. median standard deviation



شکل ۳. برچسب‌گذاری لایه‌ایِ نقش‌های ارتباطی

لایه‌ها از بالا به پایین عبارتند از تکیه، جایگاه هجا و شرایط کانونی. در این برچسب‌گذاری U و S به ترتیب به هجای تکیه‌بر و بی‌تکیه اشاره دارد، I، F و P به ترتیب هجا در جایگاه آغازی، پایانی و ماقبل پایانی است و در نهایت، PRE، ON و POS به ترتیب معرف هجا در حالت پیش‌کانون، کانون و پس‌کانون است.

پس از برچسب‌گذاری با استفاده از یکی از سه ابزارِ موجود در PENTAtrainer2 یعنی Annotation Tool، ابزار یادگیری[1] به عنوان ابزار دوم در برنامه مورد استفاده قرار گرفت تا اهداف زیروبمیِ چندنقشی (در مجموع ۱۸ مورد) را بدست آورد. ترکیب نقش‌های ارتباطی بدست آمده که در نهایت منجر به استخراج ۱۸ هدف زیروبمی متفاوت است در شکل ۴ به تصویر کشیده شده است. در هر دو تصویر ۳ و ۴ لایهٔ اول، لایهٔ نقشیِ تکیه است که شامل دو هجای تکیه‌بر (S) و غیرتکیه‌بر (U) است. لایهٔ دوم جایگاه هجا را درون کلمه نشان می‌دهد که شامل هجای آغازی (I)، پایانی (F) و ماقبل‌پایانی (P) است. لایهٔ آخر، موقعیت هجای مورد نظر را در جایگاه پیش‌کانون (PRE)، کانونی (ON) و پس‌کانون (POS) ترسیم کرده است.



شکل ٤. ترکیب نقش‌های ارتباطی بدست آمده که شامل ۱۸ ترکیب مختلف است.

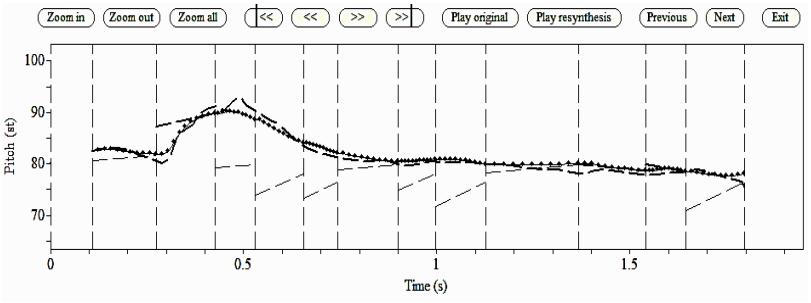
در این برچسب‌گذاری نیز U و S به ترتیب هجای تکیه‌بر و بی‌تکیه هستند، I، F و P به ترتیب به هجا در جایگاه آغازی، پایانی و ماقبل پایانی اشاره دارد و در نهایت، PRE، ON و POS به ترتیب معرف هجا در حالت پیش‌کانون، کانون و پس‌کانون است.

1. Learn tool

در ادامه، ابزار بازسازی[1] مورد استفاده قرار گرفت تا منحنیِ بازسازیِ نهایی هر یک از پاره‌گفتارها را تولید کند. لازم به ذکر است که بخش یادگیری و بازسازی به شیوهٔ مستقل از گوینده انجام گرفت. یعنی در ابتدا اهداف زیروبمی از کلیهٔ گوینده‌ها استخراج شد و سپس از بین گوینده‌ها میانگین‌گیری شد. در ادامه، از اهداف زیروبمیِ میانگین‌گیری‌شده استفاده شد تا بازسازی پاره‌گفتارهایِ کل گوینده‌ها انجام گیرد. در نهایت، ۱۵۰ پاره‌گفتار مورد نظر به همراه منحنی F_0 بازسازی‌شدهٔ آنها بر روی رایانه ذخیره شد.

نتایج آزمون عینی

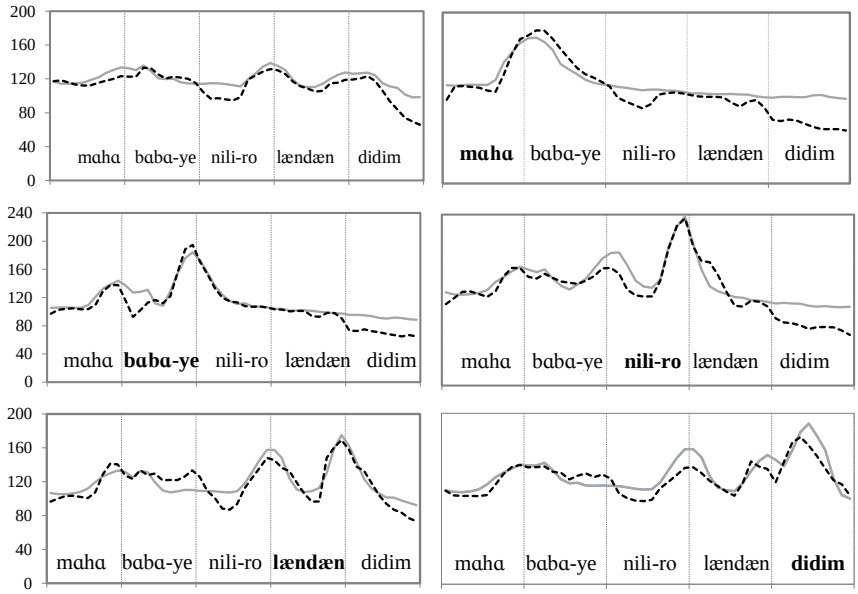
شکل ۵ نمونه‌ای از F_0 بازسازی‌شده با استفاده از ابزار بازسازی PENTAtrainer2 را نشان می‌دهد. محور افقی، زمان و محور عمودی، منحنیِ F_0 گفتار است. در شکل زیر، منحنی بریده، خطوط مورب و منحنی نقطه‌چین به ترتیب، منحنیِ F_0 اصلی، اهداف زیروبمی و منحنیِ بازسازی‌شده را نشان می‌دهد. در ادامه، شکل ۶ منحنی اصلی و بازسازی‌شده در شرایط کانونی و غیرکانونیِ کلیهٔ پاره‌گفتارها را که به طور میانگین در محور زمان بهنجارشده[2] است نمایش می‌دهد. هر یک از منحنی‌ها، میانگین بسامدپایهٔ ۲۵ پاره‌گفتار است. در شکل ۶ مرز کلمات برای هر یک مشخص شده است. درضمن، کلماتی که به صورت برجسته نشان داده شده‌اند، عناصر کانونی هستند.



شکل ۵. نمونهٔ بازسازی‌شده با استفاده از PENTAtrainer2

منحنی F_0 اصلی (خط بریده)، اهداف زیروبمی فراگیری شده (خطوط مورب) و منحنی بازسازی‌شده (منحنی نقطه‌چین) در تصویر فوق هویداست.

1. Synthesis Tool
2. mean time-normalized



شکل ٦: میانگین منحنی بسامد پایهٔ اصلی و بازسازی‌شده (هر یک میانگین ۲۵ پاره‌گفتار) که در محور زمان بهنجار شده است.

بسامد پایهٔ اصلی به صورت خط خاکستری و منحنی بازسازی‌شده به صورت بریده نشان داده شده است. آوانویسی و مرز هر یک از کلمات نیز نمایش داده شده است. کلماتی که به صورت برجسته‌اند کلمات کانونی هستند.

به منظور ارزیابی عینیِ کیفیت منحنیِ تولیدشده از خطای جذر میانگین مربعات و ضریب همبستگی استفاده شده است. این دو معیار به صورت خودکار توسط نرم‌افزار PENTAtrainer2 استخراج و ذخیره شده است. جدول ۲ نتایج حاصل از این دو معیارِ رایج یعنی خطای جذر میانگین مربعات و ضریب همبستگی را برای پاره‌گفتارهای محرک کانونی و غیرکانونی نشان می‌دهد. نتایج بدست آمده بیانگرِ میزانِ نیکویی برازشِ[1] بسیار خوب بین منحنی بازسازی‌شده و منحنی اصلی است.

جدول ۲: میانگین ضریب همبستگیِ Pearson's r و خطای جذر میانگین مربعات بین دو نوع پاره‌گفتارهایِ محرک کانونی و غیرکانونی

نوع جمله	همبستگی	RMSE
غیرکانونی	۰/۷۶	۱/۶۲
کانونی	۰/۸۶	۲/۰۱

1. goodness of fit

نتایج فوق حاکی از عملکرد بسیار مطلوب PENTAtrainer2 برروی دادگانِ کانونیِ زبان فارسی است. این یافته‌ها قابل مقایسه با یافته‌های تحقیق لی و همکاران (۲۰۱۴) و ژو و پروم ـ آن (۲۰۱۴) است که شباهت بسیار خوبِ منحنی تخمینی در مقایسه با منحنی اصلی را نشان می‌دهد.

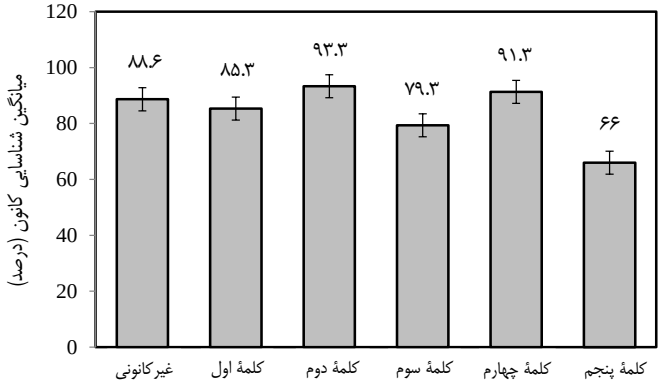
آزمون ادراکی

همان‌طور که اشاره شد برای بررسی کیفیت پاره‌گفتارهایِ با منحنی بازسازی‌شده، در کنار دو آزمون عینیِ که در بخش پیشین به آن‌ها اشاره شد، از دو آزمونِ مجزایِ ادراکی یعنی شناسایی جایگاه کانون و قضاوت میزان طبیعی بودن[1] نیز استفاده شد. در آزمون شناسایی جایگاه کانون، آزمودنی‌ها می‌بایست وجود یا عدم وجود کانون و در صورت وجود جایگاه آن را در هر یک از محرک‌ها مشخص کنند. از آنجایی که F_0 مهمترین همبستهٔ[2] کانون نوایی در فارسی است (طاهری اردلی و ژو، ۲۰۱۲؛ ابوالحسنی‌زاده، ۲۰۱۲)، در صورتی که بازسازی منحنی به‌خوبی انجام شده باشد، نتایج بدست آمده از این آزمون با آزمون شنیداری انجام شده با پاره‌گفتارهای محرکِ طبیعی قابل مقایسه است، در غیر این صورت، چنانکه میزان پاسخ‌های صحیحِ شنونده‌ها کمتر از میزان شناساییِ صحیحِ جایگاهِ کانون در پاره‌گفتارهای طبیعی باشد، می‌تواند حاکی از بازسازی نامطلوب باشد. در آزمون دوم، یعنی قضاوت پیرامون میزان طبیعی بودن، شنونده‌ها باید از بین ۱۸۰ محرک مشخص کنند که هر یک به کدام گروهِ «طبیعی» یا «بازسازی‌شده» تعلق دارند. اگر میزانِ تشخیص طبیعی بودنِ هر دو نوع محرک به یک اندازه باشد نشان از بازسازیِ با کیفیت مطلوبِ منحنی است.

نتایج آزمون ادراکی

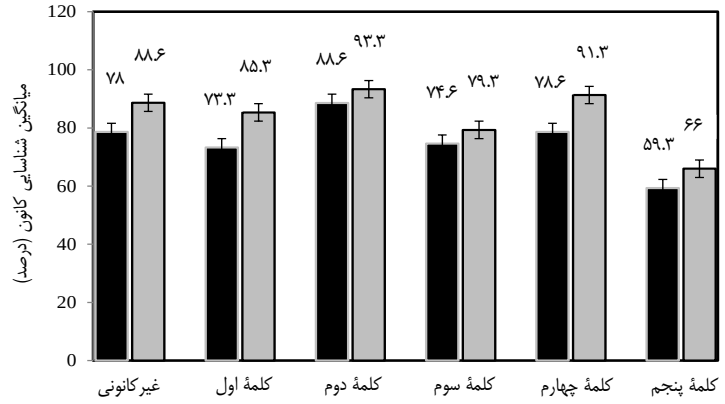
شکل ۷ نتایج حاصل از ارزیابی ذهنیِ شناساییِ کانون در آزمایش شنیداری در پاره‌گفتارهای بازسازی‌شده را نشان می‌دهد. بیشترین و کمترین میزان صحیح بازشناسی کانون، متعلق به پاره‌گفتارهایی است که به ترتیب کانون برروی کلمهٔ دوم یعنی «بابای» (۹۳.۳ درصد) و کلمهٔ پنجم «دیدیم» (۶۶ درصد) قرار دارد.

1. naturalness judgment
2. correlate



شکل ۷. درصد شناسایی صحیح کانون برروی کلمهٔ نخست تا پنجم و نیز در حالت غیرکانونی در پاره‌گفتارهای بازسازی‌شده. نوار خطا، خطای معیار را نشان می‌دهد.

به منظور مقایسه، نمودار میله‌ایِ آزمونِ شنیداریِ پاره‌گفتارهایِ اصلی از پژوهش طاهری اردلی و همکاران (۲۰۱۴) در کنار نتایج پژوهش حاضر در شکل ۸ آمده است.



شکل ۸. درصد شناسایی صحیح کانون برروی کلمهٔ نخست تا پنجم و نیز حالت غیرکانونی در دو پاره‌گفتارِ طبیعی (سیاه‌رنگ) و بازسازی‌شده (خاکستری‌رنگ). نوار خطا، خطای معیار را نشان می‌دهد.

همچنین، جدول ۳ در زیر ماتریس درهم‌ریختگیِ[1] ادراکِ جایگاه کانون را نشان می‌دهد. بیشترین درصد خطای آزمودنی‌ها در بازشناسی جایگاه کانون به شنیدن عبارت کانونیِ کلمهٔ پنجم یعنی «دیدیم» به‌عنوان عبارت غیرکانونی است. به بیان دیگر، ۲۸.۶ درصد از عبارت‌هایی که با تأکید برروی «دیدیم» بودند به اشتباه

1. confusion matrix

توسط شنوندگان به عنوان عبارات غیرکانونی قلمداد شده‌اند که با نتیجه بدست آمده در پاره‌گفتارهای طبیعی (۲۹.۳) قابل مقایسه است (قس. طاهری اردلی و همکاران، ۲۰۱۴).

جدول ۳. ماتریس درهم‌ریختگیِ ادراک کانونِ پاره‌گفتارهای بازسازی‌شده (به درصد)

اعدادی که به صورت برجسته نمایش داده شده‌اند درصد صحیح شناسایی جایگاه کانون است.

شنیده شده / حالت اصلی	غیرکانونی	کلمهٔ اول	کلمهٔ دوم	کلمهٔ سوم	کلمهٔ چهارم	کلمهٔ پنجم
غیرکانونی	**۸۸/٦**	۳/۶	۰/۰	۱/۶	۳/۶	۱/۶
کلمهٔ اول	۱۲	**۸۵/۳**	۲	۰/۰	۰/۰	۰/۰
کلمهٔ دوم	۴/۶	۲	**۹۳/۳**	۰/۰	۰/۰	۰/۰
کلمهٔ سوم	۱۳/۳	۱/۳	۳/۳	**۷۹/۳**	۲/۶	۰/۰
کلمهٔ چهارم	۸/۰	۰/۰	۰/۰	۰/۰	**۹۱/۳**	۰/۶
کلمهٔ پنجم	۲۸/۶	۰/۰	۰/۰	۰/۰	۵/۰	**٦٦/۰**

همچنین، جدول ۴ نتایج مقایسهٔ دوبه‌دوی[1] حالاتِ مختلف عبارات کانونیِ بازسازی‌شده با تعدیل بونفرونی[2] را نشان می‌دهد. تنها تفاوت معنادار بین حالت کانونی روی کلمهٔ پنجم در مقایسه با کلمهٔ دوم یعنی «بابای» و کلمهٔ چهارم «لندن» است. به بیان دیگر، میزان شناسایی صحیح کانون برای کلمهٔ پنجم به طور معناداری کمتر از شناسایی صحیح کانون برروی کلمهٔ دوم و چهارم است. این نتایج، با نتایج بدست آمده در تحقیقِ پیرامون پاره‌گفتارهای طبیعی (طاهری اردلی و همکاران، ۲۰۱۴) قابل مقایسه است. تنها تفاوت بین دو نوع پاره‌گفتار طبیعی و بازسازی‌شده، تفاوت معنادار بین کانون برروی کلمهٔ پنجم و عبارت فاقد کانون است که چنین تفاوتِ معناداری در تحقیق حاضر (عبارات بازسازی‌شده) مشاهده نشده است.

جدول ٤. نتایج مقایسهٔ دو به دو تعقیبی[3]. P < ۰/۰۵

نوع کانون (I)	نوع کانون (J)	تفاوت میانهٔ (I-J)	خطای معیار	معناداری
غیرکانونی	کلمهٔ اول	۵/۳۳۳	۵/۶۹۲	۱/۰۰۰
	کلمهٔ دوم	-۲/۶۶۹	۴/۳۵۵	۱/۰۰۰
	کلمهٔ سوم	۱۱/۳۳۳	۸/۸۳۶	۱/۰۰۰
	کلمهٔ چهارم	-۰/۶۶۸	۵/۹۱۸	۱/۰۰۰
	کلمهٔ پنجم	۲۵/۳۳۴	۶/۴۲۶	۰/۰۵۱

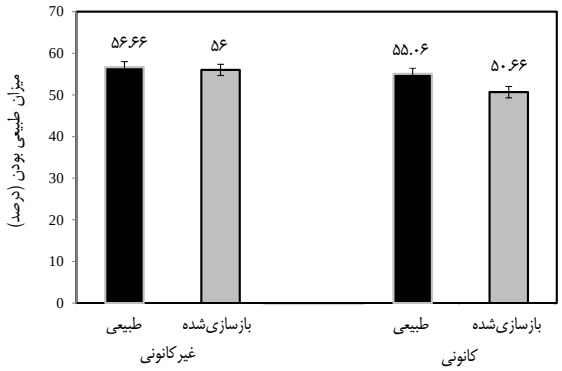
☜

1. pairwise comparison
2. Bonferroni adjustment
3. post-hoc

کلمهٔ نخست	کلمهٔ دوم	−۸/۰۰۲	۴/۰۷۴	۱/۰۰۰
	کلمهٔ سوم	۸/۰۰۰	۷/۵۳۳	۱/۰۰۰
	کلمهٔ چهارم	−۶/۰۰۱	۵/۴۸۴	۱/۰۰۰
	کلمهٔ پنجم	۲۰/۰۰۱	۸/۸۳۵	۰/۷۴۸
کلمهٔ دوم	کلمهٔ سوم	۱۴/۰۰۲	۶/۳۹۹	۰/۸۴۶
	کلمهٔ چهارم	۲/۰۰۱	۳/۱۵۱	۱/۰۰۰
	کلمهٔ پنجم	۲۸/۰۰۳*	۶/۱۱۱	**۰/۰۲۰**
کلمهٔ سوم	کلمهٔ چهارم	−۱۲/۰۰۱	۴/۴۲۳	۰/۳۵۸
	کلمهٔ پنجم	۱۴/۰۰۱	۸/۳۴۳	۱/۰۰۰
کلمهٔ چهارم	کلمهٔ پنجم	۲۶/۰۰۲*	۵/۲۰۸	**۰/۰۱۱**

نتیجهٔ جالب توجه در میزان شناسایی کانون برای پاره‌گفتارهای بازسازی‌شده است که بیشتر از درصد شناسایی جایگاه کانون در پاره‌گفتارهای طبیعی است (۸۴٪ در مقابل ۷۵/۵٪). این افزایش در میزان بازشناسی برای کلیهٔ حالت‌های کانونی و غیرکانونی بدست آمده است که در شکل ۸ قابل مشاهده است.

شکل ۹ نتایجِ دومین آزمونِ ادراکی یعنی قضاوت میزان طبیعی بودنِ هر دو نوع محرک‌های طبیعی و تولیدشده را نشان می‌دهد. اگر چه در هر دو حالت کانونی و غیرکانونی، محرک‌های بازسازی‌شده به میزان کمی درصد پایین‌تری را به خود اختصاص داده‌اند اما تفاوت معناداری در دو حالت کانونی [$F(1,9) = 2.969$, $p = 0.119$] و غیرکانونی [$F(1,9) = 0.87$, $p = 0.775$] مشاهده نشده است.



شکل ۹. میانگین (میله‌ها و اعداد روی آنها) و خطای معیار (خطوط عمودی) ارزیابی طبیعی بودن پاره‌گفتارهای بازسازی‌شده در شرایط کانونی و غیرکانونی

نتایجِ حاصل از نبودِ تفاوت معنادارِ آماری در میزان طبیعی بودنِ هر دو نوع محرک، نشان‌دهندهٔ نتیجهٔ مطلوبِ حاصل از فرایند بازسازی منحنیِ زیروبمیِ پاره‌گفتارهای کانونی با استفاده از رویکردی تولیدی- نقش‌گرا به تحلیل و بازسازیِ نوای گفتار است.

نتیجه‌گیری

مهمترین همبستهٔ نوایِ گفتار بسامدپایه است از این رو اغلب تحقیقات انجام گرفته در این زمینه در راستای تحققِ مدل‌سازیِ رایانشی این منحنی بوده است. اما در زبان فارسی، خلاءِ پردازش نوای گفتار با نگاهی زبان‌شناختی به‌ویژه در سیستم‌های تبدیل متن به گفتار به‌خوبی احساس می‌شود. دلیل چنین خلائی را می‌توان در مواردی چون دشواری ذاتی مطالعات در این حوزه، نبود تحقیقات منسجم و پیوسته برپایهٔ یک نظریهٔ واحد و همچنین نبود تعامل لازم بین زبانشناسان و متخصصان علم فناوری گفتار دانست. در مقالهٔ حاضر در اولین گام، به منظور پرکردن بخش کوچکی از این خلاء، مدل‌سازی نوای گفتارِ کانونیِ فارسی با استفاده از رویکردی جدید یعنی رمزگذاری موازی و تقریب هدف، انجام گرفت. نتایج بدست آمده نشان داد که تخمینِ نسبتاً دقیقِ F_0 در جملات خبریِ کانونیِ فارسی امکان‌پذیر است. نکتهٔ حائز اهمیت در نتایج، میزان درک کانون برای پاره‌گفتارهای بازسازی‌شده در مقایسه با نتایج حاصل از درک کانون در پاره‌گفتارهای طبیعی است که میزانِ درک کانون در جملاتِ همانندسازی‌شده در کلیهٔ حالت‌های کانونی و غیرکانونی، بهتر از همتای آن‌ها بوده است. همچنین، ارزیابی ذهنی و عینیِ منحنی‌هایِ تولیدشده حاکی از نتایج مطلوبی است که با دیگر نتایج حاصل از زبان‌هایی مانند انگلیسی، ماندرین، ژاپنی و تایلندی (لی و همکاران، ۲۰۱۴؛ ژو و پروم ـ آن، ۲۰۱۴) قابل مقایسه است. بنابراین، PENTAtrainer2 نشان داد که می‌تواند یک ابزارِ کارآمد برای همانندسازیِ نوای گفتارِ کانونی در فارسی باشد. این بستهٔ نرم‌افزاری که به صورت نیمه‌خودکار عمل می‌کند با ترکیبی از همانندسازیِ سازوکارِ تولیدی در تولید F_0، برچسب‌گذاریِ نقشی و بهینه‌سازیِ تصادفی، برای مطالعهٔ نوای گفتار زبان‌های مختلف مورد استفاده قرار می‌گیرد. اما در مقایسه با رویکردهای دیگر مانند خودواحد ـ وزنی که پیش از این در فارسی مورد استفاده قرار گرفت (طاهری اردلی و خرم، ۱۳۹۱) باید اشاره کرد که این رویکرد و استفاده از برچسب‌گذاری نواخت‌ها و فاصله‌نماها دارای موانعی است که امر مدل‌سازی را به طور کلی با مشکل مواجهه می‌کند. از آن جمله می‌توان به انجام برچسب‌گذاری به شیوهٔ دستی اشاره کرد که نه تنها انسجام و یکپارچگی لازم را ندارد بلکه بسیار زمان‌بر و هزینه‌بر است که عملاً مدل‌سازی نوای گفتار با استفاده از پایگاه‌دادگان‌های حجیم را غیرممکن می‌کند. دومین مشکل آن، خروجیِ منحنیِ F_0 به شکل ناپیوسته است که خود نیازمند پردازشِ پسینِ منحنیِ تولیدشده است. رویکرد فوجی‌ساکی شاید مطرح‌ترین و پرکاربردترین رویکرد در زمینهٔ بازسازی و مدل‌سازیِ نوای گفتار باشد که تاکنون برای زبان‌های متعددی چون انگلیسی، چینی، هندی، یونانی و غیره با میزان بالایی از موفقیت مورد استفاده قرار گرفته است. اما در این مدل به دلیل تاثیرِ مولفه‌هایِ مختلف بریکدیگر و ساختِ منحنیِ F_0 با استفاده از اثر جمعیِ مولفه‌ها، تعیین دستیِ پارامترهایِ مدل تقریباً غیرممکن است. لذا باید درصدد بود تا پارامترها به صورت

خودکار از روی منحنی اصلی استخراج شوند. اگر چه این کار با موفقیت بالایی توسط نم‌نبات و کوچاری (۱۳۸۶) انجام گرفته است اما مسئلهٔ پیش‌رو، بدست آوردن پارامترهای مورد نظر از روی متن است که پیاده‌سازی آن را در سیستم‌های تبدیل متن به گفتار فارسی با مشکل مواجه می‌کند. تیلت، مدل دیگری است که به الگوی خودواحد ـ وزنی شباهت دارد. نم‌نبات و همکاران (۱۳۸۴) در پژوهشی با استفاده از این رویکرد به مدل‌سازی نوای گفتار در فارسی و تخمین بسامدپایه از روی متن پرداختند و به میزان همبستگی حدود ۰/۶۱ برای دادهٔ آزمون خود دست پیدا کردند. این میزان از همبستگی نشان می‌دهد که به مانندِ مدل فوجی‌ساکی، بدست آوردن پارامترهای مدل از روی متن امر ساده‌ای نیست. از این رو، به باور نگارندگان این سطور، رویکردِ تولیدی ـ نقش‌گرای PENTA برخلاف رویکرد خودواحد ـ وزنی دارای قطعیت بسیار بالایی در برچسب‌های مورد استفاده است و با توجه به اینکه بسیاری از نقش‌های ارتباطی که به عنوان برچسب‌های ورودی مورد استفاده قرار گرفتند می‌توانند به شیوهٔ خودکار اعمال شوند، مدل‌سازی نوای گفتار در پایگاه‌دادگان‌های حجیم و سیستم‌های تبدیل متن به گفتار فارسی را بالقوه عملی می‌سازد. همانطور که اشاره شد از مشکلات رویکردهایی مانند فوجی‌ساکی و تیلت بدست آوردن پارامترها از روی متن است. از این رو به نظر می‌رسد رویکرد PENTA به خاطر شفافیت، انسجام و یکپارچگی آن می‌تواند به عنوان یک گزینه در مدل‌سازی نوای گفتار فارسی با استفاده از دادهٔ آموزش و آزمونِ مجزا مورد استفاده قرار گیرد. با توجه به نتایج مطلوب حاصل از پژوهش حاضر، نگارندگان در تحقیقات آتی درصددند تا این چارچوب را با تعداد نقش‌های ارتباطی فرضیِ بیشتر و با یک پایگاه‌دادگان گفتاریِ جملات خبری (طاهری اردلی و همکاران، ۱۳۹۴) که به منظور استفاده در سیستم‌های تبدیل متن به گفتار طراحی شده است، مورد آزمایش قرار دهند. یعنی، آزمودن این چارچوب با نمونه‌جملات بیشتر و با نمونه‌جملاتی که بدون نظارت مستقیم در شیوهٔ تولید، تهیه و ضبط شده‌اند. در صورت کسب نتایج قابل قبول می‌توان مدل‌سازی نوای گفتار فارسی در سیستم‌های تبدیل متن به گفتار فارسی را با استفاده از این رویکرد پیاده‌سازی کرد.

منابع

– اسلامی، محرم (۱۳۸۴)، واج‌شناسی، نظام تحلیل آهنگ زبان فارسی. تهران، انتشارات سمت.

– طاهری اردلی، مرتضی و سهیل خرم (۱۳۹۱)، «مدل‌سازی نوای گفتار در سیستم‌های سنتز گفتار فارسی»، مجموعه مقالات هشتمین همایش زبانشناسی ایران، به کوشش محمد دبیرمقدم، تهران: دانشگاه علامه طباطبایی، صفحات ۴۸۰-۴۹۲.

– طاهری اردلی، مرتضی، خرم، سهیل، عاصی، مصطفی، صامتی، حسین و محمود بی‌جن‌خان (زیر چاپ)، «طراحی و ضبط پایگاه‌دادگان گفتاری برای سیستم‌های تبدیل متن به گفتار فارسی»، دو فصلنامه علمی- پژوهشی پژوهش‌های زبان‌شناسی تطبیقی.

– نم‌نبات مجید و عباس کوچاری (۱۳۸۵)، «تخمین منحنی گام در زبان فارسی برای یک سیستم تبدیل متن به گفتار با کمک درخت کلاس‌بندی و رگرسیون»، مجموعه مقالات دوازدهمین کنفرانس ملی انجمن کامپیوتر ایران.
– نم‌نبات مجید و عباس کوچاری (۱۳۸۶)، «استخراج اتوماتیک پارامترهای مدل فوجی‌ساکی برای زبان فارسی»، مجموعه مقالات پانزدهیمن کنفرانس مهندسی برق ایران.
– Abolhasanizadeh, V., Bijankhan, M., & Gussenhoven, C. (2012), "The Persian pitch accent and its retention after the focus", Lingua, 122(13), 1380-1394.
– Bijankhan, M., Sheikhzadegan, M. J., Roohani, J., Samareh, Y., Lucas, C., & Tebyani, M. (1994), "FARSDAT-the Farsi spoken language database", Paper presented at the Proceedings of International Conference on Speech Sciences and Technology.
– Boersma, P. & D. Weenink (2001), "Praat, a system for doing phonetics by computer." Glot international (5): 341–345.
– Chen, Y., Guion-Anderson, S., & Xu, Y. (2012), "Post-Focus compression in second language Mandarin". Speech Prosody 2012, Shanghai.
– Ferguson, C. (1957), "Word stress in Persian". Language 33, 123-135.
– Fujisaki, H. & K. Hirose (1984), "Analysis of voice fundamental frequency contours for declarative sentences of Japanese." Journal of the Acoustical Society of America 5(4): 233-242.
– Fujisaki, H. & S. Nagashima (1969), "A model for the synthesis of pitch contours of connected speech." Annual Report of the Engineering Research Institute 28: 53-60.
– Jun, S.-A. (2005), Prosodic Typology: The Phonology of Intonation and Phrasing. Oxford: Oxford University Press.
– Kahnemuyipour, A. (2003), "Syntactic categories and Persian stress." Natural Language & Linguistic Theory, 21(2), 333-379.
– Ladd, R. (2008). Intonational Phonology. Cambridge, Cambridge University Press.
– Lazard, G. (1957), Grammaire du Persan Contemporain. Klincksieck, Paris, New Edition published by Peeters, Paris, 2006.
– Lee, A., Xu, Y., & Prom-on, S. (2014), "Modeling Japanese F0 contours using the PENTAtrainers and AMtrainer". Fourth International Symposium on Tonal Aspects of Languages. Nijmegen, Netherlands.
– Mahjani, B. (2003), An Instrumental Study of Prosodic Features and Intonation in Modern Farsi (Persian), M.Sc. thesis, retrieved from: http://www.ling.ed.ac.uk/teaching/postgrad/mscslp/archive/dissertations/2002-3/behzad mahjani.pdf.
– Pierrehumbert, J. B. (1980), The phonology and phonetics of English intonation. (PhD dissertation), Massachusetts Institute of Technology.
– Prom-on, S., Xu, Y., Thipakorn, B. (2009), "Modeling tone and intonation in Mandarin and English as a process of target approximation." The Journal of the Acoustical Society of America 125(1): 405–424.
– Sadat-Tehrani, N. (2007), The Intonational Grammar of Persian. (PhD dissertation), University of Manitoba, Manitoba.
– Silverman, K. E., Beckman, M. E., Pitrelli, J. F., Ostendorf, M., Wightman, C. W., Price, P., & Hirschberg, J. (1992), "TOBI: a standard for labeling English prosody". In The Second International Conference on Spoken Language Processing, ICSLP 1992, Banff, Alberta, Canada, October 13-16.

- Taheri-Ardali, M., Rahmani, H., & Xu, Y. (2014), "The perception of prosodic focus in Persian". Speech Prosody 2014. Dublin: 515-519.
- Taheri-Ardali, M. & Y. Xu (2012), "Phonetic realization of prosodic focus in Persian". Speech Prosody 2012, Shanghai.
- Taylor, P. (1992), A Phonetic Model of English Intonation. (PhD dissertation), University of Edinburgh, Edinburgh.
- Taylor, P. (2000), "Analysis and synthesis of intonation using the tilt model." Journal of the Acoustical Society of America 107(4): 1697–1714.
- Xu, Y. (2005), "Speech melody as articulatorily implemented communicative functions." Speech Communication 46(3–4): 220–251.
- Xu, Y. & S. Prom-on (2014), "Toward invariant functional representations of variable surface fundamental frequency contours: Synthesizing speech melody via model-based stochastic learning." Speech Communication 57: 181–208.
- Xu, Y. & E. Q. Wang (2001), "Pitch targets and their realization: Evidence from Mandarin Chinese." Speech Communication 33: 319-337.