مقدمه

در ادامه تلاش برای برقراری ارتباط طبیعی بین انسان و ماشین و پس از ارائه سیستم بازشناسی گفتار پیوسته برای زبان فارسی[1]، در این مقاله مراحل طرح و پیاده‌سازی یک سیستم نمونهٔ تبدیل متن به گفتار طبیعی برای زبان فارسی ارائه می‌شود. بر خلاف سیستم‌های پاسخگوی صوتی (Voice Response Systems)، TTS قابلیت ادای طبیعی جملات جدید را دارد [2]. شکل (۱) شمای بلوکی کلی TTS طراحی شده در این تحقیق را نشان می‌دهد.

در بلوک NLP جملات ورودی به فهرستی از کلمات تبدیل می‌گردند و در این راستا، اعداد، حروف اختصاری (Abbreviations) و ... نیز به معادل کامل متنی خود تبدیل می‌شوند. افزون بر این، روی کلمات جمله تحلیل ریشه‌ای (Morphological) انجام و مشخصه دستوری هر کلمه (POS) تعیین می‌گردد. در مورد لزوم پردازش نحوی (Syntactic) متن در بلوک NLP نیز باید گفت که تبدیل صحیح متن فارسی به رشته واجی با تعیین مشخصه دستوری کلمات (به عنوان مثال از طریق بررسی در قالب درختوارهٔ دستوری (Syntax tree) [3]) تسهیل می‌شود. نوای گفتار فارسی نیز بستگی زیادی به ساختار دستوری جمله دارد [۴]. مرحلهٔ بعدی کار در بخش NLP استخراج صورت واجی متن است. شاید بدیهی‌ترین راه انجام این کار، استفاده از مجموعه قواعد تبدیل حرف به صدا به نظر برسد که حروف متن را به رشته‌ای از واج‌ها تبدیل کند. این راه‌حل برای زبان‌هایی همچون زبان اسپانیایی که رابطهٔ نزدیکی بین صورت نوشتاری و گفتاری آنها وجود دارد، می‌تواند مفید باشد، ولی برای زبان‌هایی چون فارسی و حتی انگلیسی که تناظر مستقیمی بین حروف و واج‌ها وجود ندارد، باید به دنبال راه‌حل بهتری بود[5]. به عنوان نمونه در زبان فارسی می‌توان به مشکلات زیر در زمینهٔ استخراج صورت واجی متن اشاره کرد:

الف) وجود یک حرف برای چند واج (مانند تلفظ‌های متفاوت "و" در کلمات "تو"، "او" و "وزن")

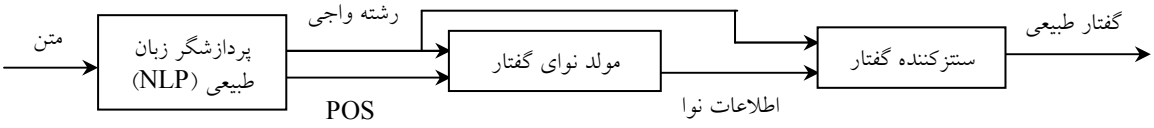
ب) استفاده از حرفی که خوانده نمی‌شود (مانند "و" در کلمه "خواستن")

پ) نوشته نشدن واکه‌های کوتاه (نگذاشتن اعراب)

ت) نوشته نشدن کسرهٔ اضافه در ترکیب‌های وصفی و اضافی

در خصوص بلوک مولد نوای گفتار نیز باید گفت که "زیر و بمی" یا تغییرات فرکانس گام (Pitch)[6-8]، "درنگ" (Duration) [9-11]، "شدت" (Intensity) [12-13] و "درنگ"(Pause) [14] چهار عنصر نوایی گفتار هستند که معمولاً در سطوح مختلف اعم از هجا (Syllable) ، واژه و یا جمله اثر خود را نشان می‌دهند. اعمال اطلاعات نوا به سیستم سنتزکننده، نقش بسزایی در تولید گفتار طبیعی در زبان‌های مختلف دارد [15-21]. برای سنتز گفتار نیز دو روش کلی مبتنی بر قاعده (Rule-based) ، که در آن پارامترهای مشخصهٔ گفتار در هر بازهٔ زمانی توسط مجموعه‌ای از قواعد تولید می‌شوند و نیز، روش اتصال قطعات گفتار (Concatenation) که در آن واحدهای از پیش ذخیره شدهٔ صوتی برای تولید عبارتی دلخواه در کنار هم چیده می‌شوند، ارائه شده است [22].

ساختار مقاله حاضر بدین‌صورت است که در بخش دوم آن جزئیات کار در NLP و نوآوری های مربوط که از مهمترین آنها می‌توان به حل مشکل کسرهٔ اضافه و رفع ابهام در کلمات با نوشتار یکسان مبتنی بر تحلیل نحوی اشاره کرد، آورده می شود. در بخش سوم چگونگی آماده‌سازی خودکار اطلاعات آموزشی برای سیستم مولد نوای گفتار که در جریان آن تقطیع و نامگذاری دقیق قطعات گفتاری در سطح واج صورت می‌پذیرد، ارائه می گردد. در بخش چهارم ساختار سیستم مولد نوای گفتار که یک شبکه عصبی بازگشتی با ۱۳۰ نرون در لایهٔ اول، ۱۲ نرون در لایهٔ خروجی و در کل دارای ۲۸۹ نرون در چهار لایه است، معرفی خواهد گردید. در بخش پنجم چگونگی سنتز گفتار به روش HNM که در این طرح انتخاب و پیاده‌سازی شده و امروزه از طرح‌های موفق سنتز تلقی می گردد و AT&T نیز آن را در سیستم NEXT-GEN



شکل۱ شمای بلوکی سیستم TTS فارسی

خود بکار گرفته است [23]، مورد بحث قرار گرفته و نوآوری‌های مربوط در آن آورده خواهد شد. در بخش ششم نیز نتایج عملکرد کلی سیستم مورد ارزیابی قرار می‌گیرد.

پردازشگر زبان طبیعی

در پردازش زبان‌طبیعی برای نمایش‌معرفت (Knowledge) دو روش کلی وجود دارد: صرفی (Declarative) و روندی (Procedural). در روش اول معرفت به صورت حقایقی (Facts) مشخص بیان می‌شود، در حالی که در روش دوم به کمک مجموعه‌ای از قواعد، معرفت لازم بدست می‌آید.

در این پردازشگر برای دسترسی به معرفت از هر دو روش استفاده می‌شود و در این راستا پایگاه داده‌ای شامل کلمات، ساختار واجی آنها و نوع کلمه (از لحاظ دستوری) و با امکاناتی برای ورود، جستجو، حذف، نمایش کلی و مرتب کردن تمام کلمات در نظر گرفته شده است. اطلاعات هر کلمه نیز در حوزه‌های زیر وارد پایگاه داده می‌شود:

الف) Sword که از نوع Cstring بوده و حاوی کلمهٔ فارسی مورد نظر است.

ب) Sphonetic که از نوع Cstring بوده و حاوی ساختار واجی کلمه است.

پ) nType که از نوع Integer بوده و نوع دستوری کلمه (انواع فعل، صفت، اسم، مصدر، صوت، ضمیر، قید و حرف اضافه) را نمایش می‌دهد.

در این پردازشگر توابع مختلفی در نظر گرفته شده که دیگر نیازی به ورود شکل‌های مختلف یک کلمه در پایگاه داده نباشد. به عنوان نمونه برای افعال تنها "بن ماضی" و "بن مضارع" فعل در پایگاه وارد می‌شود و انواع افعال گذشته (ساده، استمراری، التزامی، نقلی ساده، نقلی استمراری و بعید)، حال (ساده، استمراری و التزامی)، آینده (ساده و استمراری)، امر، نهی، منفی و مجهول و نیز ۶ صورت مختلف صرفی آنها (اول شخص مفرد تا سوم شخص جمع) در متن تشخیص داده شده و ساختار آوایی آنها به صورت خودکار ساخته می‌شود [۲۴]. فهرست این توابع تشخیص و ساخت واجی ترکیب‌های مربوط نیز چنین است:

الف- فعل

ب- اسم و صفت جمع

پ- صفت تفضیلی و عالی

ت- اسم و صفت نکره

ث- ضمیر متصل

ج- اعداد

چ- مصدر ساده و مرکب

ح- کسرهٔ اضافه ("ی" افزوده) در ترکیب‌های وصفی و اضافی

در زبان فارسی برای تولید جملات قاعده‌های ساخت مختلفی وجود دارد که در این پردازشگر در نظر گرفته شده‌اند[3]. نمونه‌هایی از این قواعد عبارتند از:

الف) جمله ← گروه اسمی (اسم + صفت تفضیلی + "ی") + گروه فعل ربطی (حرف اضافه + گروه اسمی + فعل ربطی).

ب) جمله ← گروه اسمی (ضمیر اشاره + اسم) + گروه فعل لازم (حرف اضافه + گروه اسمی + فعل لازم).

پ)جمله ← گروه اسمی (صفت عالی + اسم +

(صفت) + "ی") + گروه فعل متعدی (حرف اضافه + گروه اسمی + گروه اسمی + را + فعل متعدی).

یادآوری می‌شود که چون قید می‌تواند در موقعیت‌های مختلف گزاره و یا قبل از نهاد قرار گیرد، در الگوهای سازه‌ای فوق آورده نشده است، ولی در الگوریتم پیاده شده در تمام موقعیت‌های ممکن در نظر گرفته شده است. نحوهٔ کار پردازشگر بدین صورت است که سیستم یک جمله کامل (که ترجیحاً با علامت گذاری خاتمه یافته است) را دریافت می‌نماید و سپس در دو مرحله پردازش جمله را انجام می‌دهد:

الف) کلمات را که با فاصله از یکدیگر جدا شده‌اند ، با استفاده از پایگاه داده و توابع تشخیص کلمات ترکیبی شناسایی کرده و اطلاعات لازم را برای پیمایشگر فراهم می آورد. در این مرحله شکل واجی کلمات نیز ساخته می‌شود.

ب) با استفاده از اطلاعات فوق (بند الف) پیمایشگر اقدام به شناسایی نوع جمله و تعیین مشخصه دستوری کلمات و شکل واجی در ترکیب‌های اضافی و وصفی می‌نماید.

در خصوص تشریحگر(Parser) NLP نیز با توجه به ساختار زبان فارسی، مناسب‌ترین روش برای مدل‌سازی آن استفاده از دو روش بالا به پایین(Top-down) و پایین به بالا (Bottom-up) به صورت توأم می‌باشد. بدین‌ترتیب که تشریحگر مورد استفاده در سیستم بر اساس روش بالا به پایین و مستقل از متن (Context-free) است ولی در مواردی منطق برنامه به گونه‌ای تغییر داده می‌شود که پیمایشگر بطور مقطعی روش پایین به بالا را بخود بگیرد و سپس به روند اصلی خود (بالا به پایین) برگردد.

خروجی بخش NLP، افزون بر رشتهٔ واجی متن، مشخصه دستوری (POS) هر یک از کلمات/ عبارات متن است که در این سیستم، ۴۱ مشخصه گزارش می‌شود (جدول ۱).

گفتنی است که در NLP انتخاب تعداد مشخصه دستوری کلمات منطبق بر نیازمندی بخش مولد نوای گفتار است.

در خصوص رفع ابهام در کلمات با نوشتار یکسان نیز از مجموع ۱۱۰ کلمه دارای این ابهام در زبان فارسی، ۵۲ کلمه رفع ابهام شده‌اند (بدون نیاز به تحلیل معنایی).

جدول ۱ مشخصه دستوری (POS) کلمات (خروجی بخش NLP)

مشخصه دستوری کلمه (POS)	کد عددی تخصیص‌یافته	مشخصه دستوری کلمه (POS)	کد عددی تخصیص‌یافته	مشخصه دستوری کلمه (POS)	کد عددی تخصیص‌یافته
فعل لازم	۱	مصدر ساده	۱۵	قید مقدار	۲۹
فعل متعدی معلوم	۲	مصدر مرکب	۱٦	سایر قیود	۳۰
فعل متعدی مجهول	۳	مصدر جمع	۱۷	حرف اضافه مفرد (بدون کسره)	۳۱
فعل معین	٤	صوت	۱۸	حرف اضافه مرکب (بدون کسره)	۳۲
فعل امر	۵	ضمیر منفصل	۱۹	حرف اضافه مفرد (کسره‌پذیر)	۳۳
فعل نهی	٦	ضمیر متصل	۲۰	حرف اضافه مرکب (کسره‌پذیر)	۳٤
صفت	۷	ضمیر اشاره	۲۱	حرف ربط مفرد	۳۵
صفت نکره	۸	مبهمات	۲۲	حرف ربط مرکب	۳٦
صفت تفضیلی	۹	ادوات پرسش	۲۳	علامت مفعول بی‌واسطه (را)	۳۷
صفت عالی	۱۰	عدد	۲٤	حرف اضافه همراه با مفعول (مانند مرا)	۳۸
اسم جمع	۱۱	وجه وصفی	۲۵	فعل رابط	۳۹
اسم جمع نکره	۱۲	قید زمان	۲٦	موصوف	٤۰
اسم	۱۳	قید مکان	۲۷	مضاف	٤۱
اسم نکره	۱٤	قید حالت	۲۸		

آماده‌سازی خودکار اطلاعات آموزشی برای سیستم مولد نوای گفتار

در این قسمت روش تقطیع سیگنال گفتار در سطح رویداد (Event) [25] و نیز چگونگی تعیین واکداری، بی‌واکی و سکوت این رویدادها آورده می‌شود. در ادامه چگونگی برچسب‌زنی(Labeling) سیگنال گفتار در سطوح مختلف جمله (کلمه، هجا و واج) بیان خواهد گردید. بدین‌ترتیب امکان استخراج خودکار اطلاعات نوا فراهم آمده و داده‌های آموزشی برای سیستم مولد نوای گفتار تأمین می‌شوند.

تعیین واکداری / بی‌واکی / سکوت. در این عملیات فرکانس نمونه‌برداری ۱۰ کیلوهرتز، نمونه‌ها ۱۶ بیتی و قطعات ۲۵ میلی ثانیه‌ای هستند. برای تعیین واکداری / بی‌واکی قطعات در ابتدا، اولین ضریب انعکاس (Reflection coefficient)، سیگنال مانده (Residue)، انرژی و ضرایبAR (Auto-Regressive) محاسبه می‌شوند. مقدار انرژی نیمه‌های اول و دوم فریم جاری را نیز بدست آورده (آنها را energy1 و energy2 نامیده) و مقدار شاخص انرژی را به صورت زیر تعیین می‌کنیم [25]:

$$\text{energy index} = \sqrt{\text{energy1} \times \text{energy2}} \qquad (۱)$$

درصورتی‌که مقدار اولین ضریب انعکاس از آستانه ۰/۴ و شاخص انرژی یاد شده (با فرض ضرب سیگنال در ضریب ((14000/max (abs(signal))) از آستانه $3\text{x}10^{7}$ به صورت همزمان فراتر روند، قطعه را به صورت تجربی واکدار در نظر می‌گیریم (و برای این قطعات فرکانس گام، دامنه و فرکانس فرمنت‌ها را محاسبه می‌کنیم). در غیر این صورت قطعات بی‌واک/ سکوت هستند که برای تشخیص نوع آنها نیز ابتدا تابع حجم (Volume function) در بازهٔ فرکانسی ۳۰۰ هرتز تا ۵ کیلوهرتز چنین تعیین می‌شود:

$$V(i) = \frac{1}{N_i}\sqrt{\sum_{m=17}^{256}\left|H_i(e^{j\pi m/256})\right|} \qquad (۲)$$

که i اندیس فریم جاری، N_i تعداد نمونه‌های فریم و $H_i(z)$ تابع انتقال با رابطه زیر است:

$$H_i(z) = \frac{G(i)}{1 + a_1 z^{-1} + \ldots + a_{13} z^{-13}} \qquad (۳)$$

G(i) نیزعبارت بهره بوده و از رابطهٔ زیر بدست می‌آید:

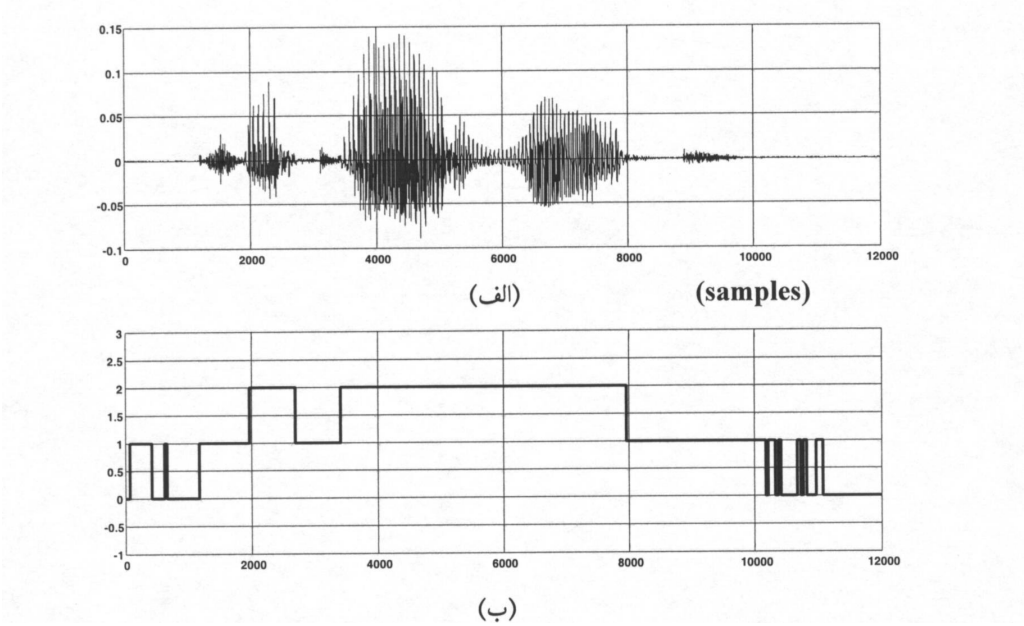
$$G(i) = \sqrt{\sum_{n=<250>} r^2(n)} \qquad (۴)$$

r(n) هم سیگنال مانده فریم جاری است. با داشتن مقدار بهره و ضرایب AR، توان چند میلی ثانیه ابتدایی سیگنال (که سکوت است) را بدست آورده و مقدار میانگین(m) و انحراف معیار(σ) آن را تعیین می‌کنیم. در این راستا، آستانه‌ای به صورت زیر تعیین و در نظر گرفته می‌شود:

$$T_{U/S} = m + K.\sigma \qquad (۵)$$

مقدار K نیز به صورت تجربی و در جریان ادای ۱۰۰ جمله (با SNR=30dB) برابر ۲ در نظر گرفته شده است. در صورتی که $20\log_{10} V(i)$ فریم مورد نظر بیش از $T_{U/S}$ باشد، آن را بی‌واک و در غیر این صورت آن را به عنوان سکوت در نظر می‌گیریم. البته بدیهی است که در جریان این پیش پردازش ساده، امکان بروز اشتباه، بویژه در تشخیص قطعات بیواک و سکوت، وجود دارد. اما با توجه به مراحل پردازش تکمیلی ارائه شده در این مقاله خطای موجود به حداقل می رسد. شکل (۲) نوع قطعات (V/U/S) را برای عبارت "چطور بود" نشان می‌دهد (V/U/S متناظر با سطوح 2/1/0) فرکانس‌های گام و فرمنت نیز با توجه به امکانات ارائه شده در مرجع [25] تعیین شده‌اند.

تقطیع رویدادها. گفتار را در سطح جزئی تحت عنوان رویداد تقطیع کرده تا پس از تعیین ارتباط بین رویدادها، سیگنال در سطح واج قطعه‌بندی شود. برای تعیین مرز این رویدادها نیز از بررسی تغییرات



شکل ۲ تعیینV/U/S قطعه گفتاری، الف) سیگنال زمانی، ب) نوع قطعه

جدول ۲ مقادیر پارامترهای بکار رفته در تعیین امتیاز دسته‌های آکوستیکی همخوان واکدار ، شبه واکه و قطعات با فرکانس فرمنت خیلی کم (بعنوان دسته کمکی)

نام اختصاری امتیاز	T_{low}	T_{up}	HFV			LFV			نوع دسته
			B	A	G	B	A	G	
VCS	۸	۱۸	۲۵۵	۵۲	۱	۵۱	۱	۱	همخوان واکدار
VBS	۱۰	۳۰	۲۵۵	۳٤	۱	۳۳	۱	۱	قطعه با فرکانس فرمنت خیلی کم
MS (Murmur Score)	۴	۱۲	۵۰	۲۱	۱	۲۰	۱	۱	شبه واکه

مشخصات زیر به عنوان سه روش مکمل استفاده شده است:

الف) وابستگی طیفی کوتاه مدت سیگنال

ب) مشخصات آکوستیکی

پ) دامنه و فرکانس فرمنت‌ها

البته روش اول در تعیین مرز واج‌های انفجاری با مشکل مواجه است. در روش دوم نیز اساس کار بر تقسیم‌بندی گفتار برحسب انواع رسا (Sonorant)، واکه (Vowel) همخوان واکدار (Voiced consonant)، خیشومی (Nasal)، شبه واکه (Semivowel) و سایشی واکدار(Voiced fricative) است. در این روش سیگنال گفتار برحسب شش نوع یاد شده دسته‌بندی و امتیازدهی می‌شود. بدین‌ترتیب که تابع حجم رابطهٔ (۲) با حدود سیگمای A و B (مبین فرکانس قطع پایین و بالای یک فیلتر میان‌گذر) در نظر گرفته می‌شود. $H_i(z)$ نیز همان است که در رابطهٔ (۳) ارائه شد. سپس توابع حجم فرکانس پایین (Low Frequency Volume function) LFV و فرکانس بالا (High Frequency function) HFV Volume به ازای مقادیر متفاوتی از G، A و B محاسبه و نسبت آنها تشکیل می شود:

$$R(i) = \frac{LFV(i)}{HFV(i)} \qquad (۶)$$

پس از اعمال یک فیلتر میانه مرتبهٔ ۵ به R (i)، امتیاز برخی از دسته‌ها، به عنوان نمونه، چنین محاسبه می‌شود:

$$Score(i) = \begin{cases} 1 & ; \ R(i) \geq T_{up} \\ 0 & ; \ R(i) < T_{low} \\ \dfrac{R(i) - T_{low}}{T_{up} - T_{low}} & ; \ T_{low} \leq R(i) < T_{up} \end{cases} \qquad (۷)$$

مقادیر بکار رفته برای محاسبهٔ امتیاز یاد شده برای اصوات همخوان واکدار، قطعات با فرکانس فرمنت خیلی پایین (Voice bar) (یعنی حدود ۱۵۰ هرتز که برای محاسبه امتیاز شبه واکه‌ها استفاده می‌شود) و شبه واکه براساس روند یاد شده به عنوان نمونه در جدول (۲) آورده شده است.

البته در مورد شبه واکه ها، امتیاز مربوط(SVS(i)) پس از تعیین(VCS(i)،VBS(i)وMS(i) چنین محاسبه می شود:

$$SVS(i) = (1 - MS(i))(1 - VBS(i))VCS(i) \qquad (۸)$$

برای دسته‌های دیگر نیز امتیاز با قدری تغییر در روابط قابل محاسبه است [۲۶و25]. شکل (۳) نتایج این امتیازبندی را برای قطعات غیرواکه گفتار موضوع شکل (۲) نشان می‌دهد.

روش سوم نیز در مواقعی که دو روش قبلی نتوانند مرز بین واج‌های واکدار را (بهنگام ادای گفتار پیوسته سریع) آشکار کنند، بکار گرفته می‌شود. گفتنی است که بدین ترتیب عملکرد این سیستم مانند روش دو مرحله‌ای ارائه شده در مرجع [27] از لحاظ دقت قابل مقایسه با تقطیع دستی قطعات می‌باشد.

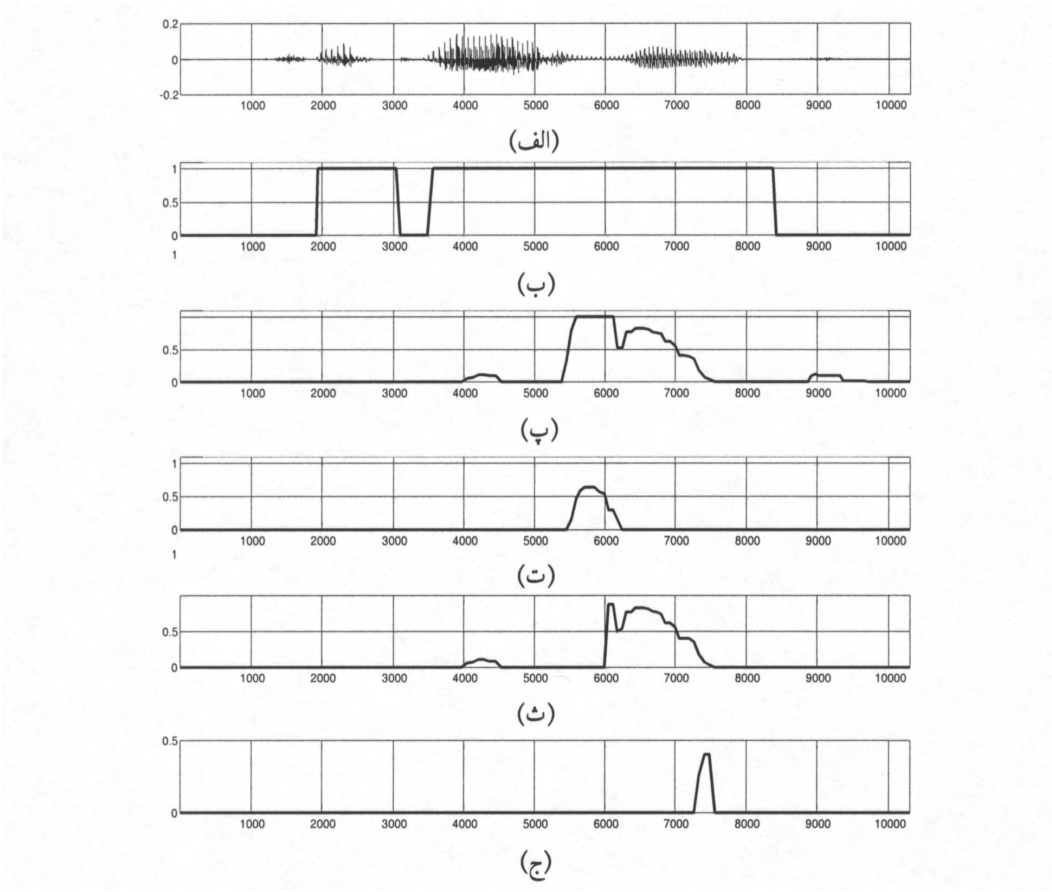
نامگذاری خودکار قطعات مختلف گفتار. برای تشخیص واج‌ها و نیز تعیین مرزهای هجاها و کلمات در جملات ادا شده، مراحل کاری زیر انجام شده است:

الف) تشخیص واج‌های واکدار

ب) تشخیص واج‌های بی‌واک و سکوت

پ) نامگذاری واج‌های واکدار با توجه به محل واکه‌ها

ت) نامگذاری همهٔ واج‌های جمله



شکل ۳ امتیازبندی آکوستیکی قطعات گفتاری، الف) قطعه گفتاری، ب) امتیاز رسا، پ) امتیاز همخوان واکدار، ت) امتیاز خیشومی، ث) امتیاز شبه واکه، ج) امتیاز سایشی واکدار

برای تشخیص واج‌های واکدار از یک شبکهٔ عصبی پرسپترون چند لایهٔ (Multi-Layer MLP Perceptron) با ۹ گره در لایه ورودی (شامل اطلاعات فرکانسی سه فرمنت اول قطعه جاری و دو قطعه مجاور آن)، دو لایه میانی به ترتیب با ۲۰ و ۳۰ گره و ۲۹ گره در لایهٔ خروجی (به تعداد واج‌های زبان فارسی) استفاده شده است. البته قطعاتی از واج‌های واکدار به خاطر مجاورت با قطعات بی‌واک/ سکوت ممکن است رفتار مشابه آنها را از خود نشان دهند. امتیازدهی این قطعات به روش امتیازدهی به واج‌های بی‌واک/ سکوت خواهد بود.

برای تشخیص واج‌های بی‌واک/ سکوت نیز از MLP دیگری با ۹ گره در لایه ورودی (شامل انرژی سیگنال در فرکانس‌های بیش از ۲KHz، انرژی سیگنال در فرکانس‌های ۵KHz-۰ و نرخ عبور از صفر مربوط به قطعه جاری و دو قطعه مجاور آن)، دو لایه میانی به ترتیب با ۲۰ و ۳۰ گره و نیز ۳۰ گره در لایهٔ خروجی (به تعداد واج‌های زبان فارسی و سکوت) استفاده شده است.

با استفاده از امتیازدهی شبکه عصبی یاد شده و نیز با توجه به متن ورودی، نامگذاری واج‌های واکدار صورت می‌گیرد. برای جلوگیری از انتشار خطا نیز ابتدا محل یک قطعه از هر واکه تشخیص داده می‌شود. بدین‌ترتیب خطای نامگذاری بین دو واکه محدود می‌شود. برای تعیین محل یک قطعه از هر واکه نیز در منحنی بهره (رابطه ۴) به دنبال بیشینه‌های محلی می‌گردیم. تعداد این بیشینه‌ها معمولاً بیش از تعداد واکه‌هاست. برای حذف بیشینه‌های اضافی نیز ۲۵ معیار بکار گرفته شده‌اند که به نمونه‌هایی از آنها در زیر اشاره می‌کنیم (که به صورت تکی یا ترکیبی اعمال می شوند):

الف) نسبت نرخ عبور از صفر نرمالیزه شده به طول فریم، کمتر از $\frac{1}{3}$ باشد.

ب) مقدار انرژی مؤلفه‌های فرکانسی کمتر از ۲/۵ کیلوهرتز، بیشتر از مؤلفه‌های فرکانسی ۲/۵ تا ۵ کیلوهرتز باشد.

پ) برای دو قطعه مجاور که واکه تشخیص داده شده‌اند و فاصلهٔ آنها کمتر از ۱۰۰ میلی ثانیه است، همبستگی طیفی کمتر از ۰/۸۵ باشد.

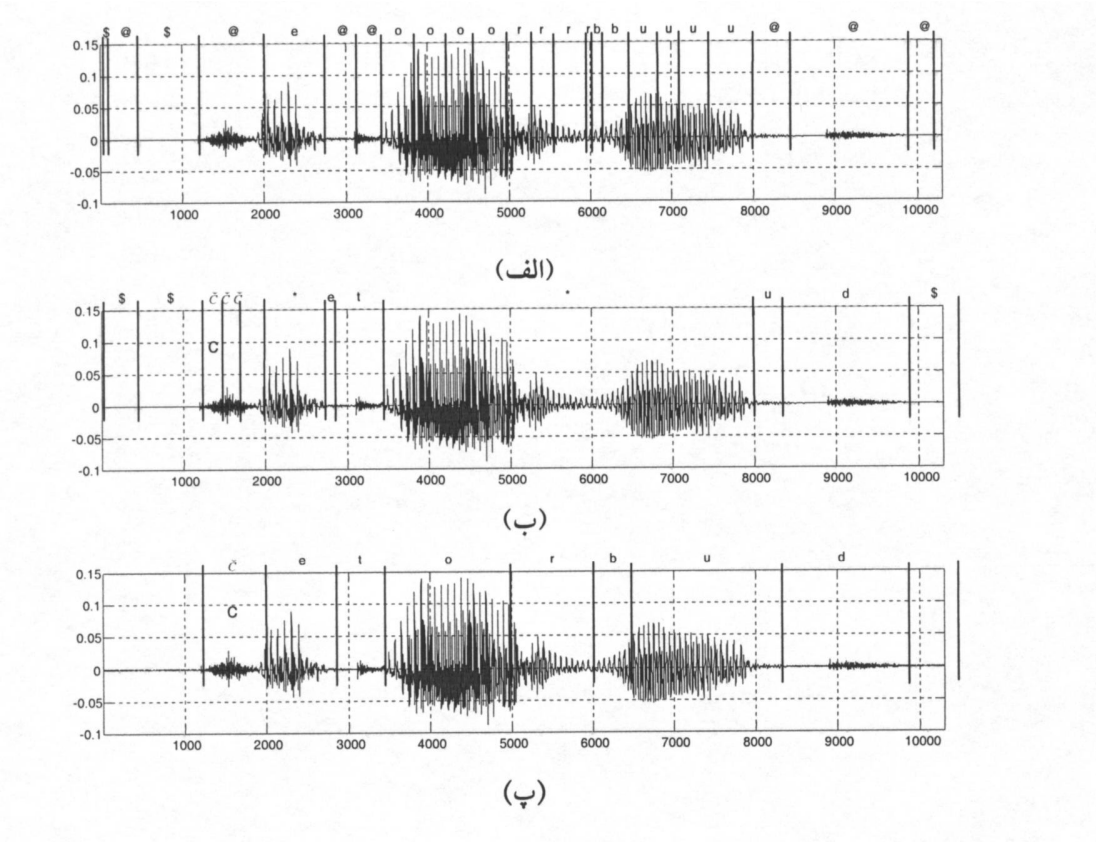
دقت این الگوریتم ۹۷/۲ درصد است و برای رسیدن به دقت ۱۰۰ درصد، روش دستی اعمال شده است. در این راستا نتایج نامگذاری قطعات واکدار، بی‌واک و واجی قطعات در شکل (۴) نشان داده شده است.

برای تعیین مرز هجاها نیز با توجه به‌حالت‌های مختلف هجایی در زبان فارسی (CV، CVC و CVCC)، با پیدا کردن محل واکه‌ها و در نظر گرفتن واج قبل از آن می‌توان به راحتی مرزهای هجا را مشخص نمود. با توجه به تقطیع و نامگذاری خودکار قطعات می‌توان پارامترهای نوایی مربوط (الگوی فرکانس گام، دیرش، شدت و درنگ) را تعیین نمود. بدین‌ترتیب داده‌های آموزشی به صورت خودکار برای سیستم مولد نوای گفتار فراهم می‌آید.

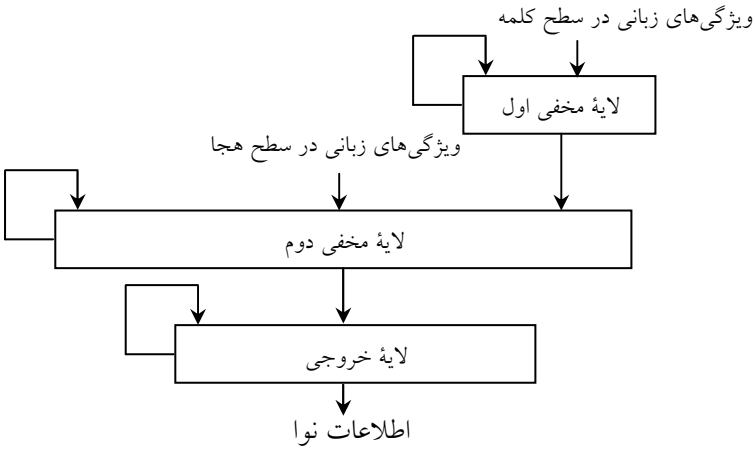
سیستم مولد نوای گفتار

روش‌های تولید اطلاعات نوا را می‌توان به دو دستهٔ کلی مبتنی بر قاعده [15,28-35] و مبتنی بر آموزش (که این کار اغلب توسط مدل‌های آماری [36-43] یا شبکه‌های عصبی[7-9,14,-44-47] و یا مدل‌های آمیختار [48] انجام می‌پذیرد) تقسیم نمود. در سیستم حاضر از یک شبکه عصبی بازگشتی (RNN) چهار لایه برای تولید نوا استفاده شده است [19-49]. شکل(۵) ساختار بلوکی این RNN را نشان می‌دهد.

ویژگی‌های ورودی در سطح کلمه عبارتند از: مشخصهٔ دستوری کلمه جاری و بعدی (با توجه به ۴۱ مشخصه ارائه شده در جدول ۱، که توسط بخش NLP تأمین می‌شود)، طول کلمه جاری و بعدی (برحسب تعداد هجا) و نوع علامت (″نقطه″، ″ویرگول″، ″علامت سؤال″، ″علامت تعجب″ و″نقل قول″) پس از کلمه جاری.



شکل ٤ نامگذاری قطعات گفتاری، الف) قطعات واکدار ($ و @: قطعات نامگذاری نشده)، ب) قطعات بی‌واک ($: سکوت و *: قطعات واکدار نامگذاری شده)، پ) نامگذاری واجی قطعات



شکل ٥ ساختار بلوکی RNN مولد اطلاعات نوای گفتار

با توجه به ۴۱ نوع مشخصهٔ دستوری و در نظر گرفتن حداکثر ۵ هجا در هر کلمه و با فرض کدینگ باینری در ورودی شبکه، تعداد نرون‌های مبین ویژگی‌های زبانی در سطح کلمه، ۹۷ خواهد بود. ویژگی‌های ورودی در سطح هجا نیز عبارتند از: نوع آغازین (همخوان اولیه) هجای جاری و بعدی (با توجه به دسته‌بندی همخوان‌ها

و به نحوهٔ ادا [50] و بر اساس جدول ۳)، نوع واکه هجای جاری و بعدی (با توجه به ۶ واکه در زبان فارسی)، نوع همخوان دوم و سوم هجای جاری (با توجه به ساختار CV، CVC و CVCC برای هجاها در زبان فارسی) و موقعیت هجای جاری در کلمه (با در نظر گرفتن چهار حالت "تک هجایی"، "هجای ابتدایی"، "هجای میانی" و "هجای انتهایی"). با توجه به امکان عدم حضور همخوان دوم و/ یا سوم در هجا و با فرض کدینگ باینری در ورودی این لایه از شبکه نیز تعداد نرون‌های مبین ویژگی‌های زبانی در سطح هجا، ۴۲ خواهد بود.

جدول ۳ دسته‌بندی همخوان‌های زبان فارسی (به عنوان اطلاعات ورودی لایهٔ مخفی دوم RNN)

شماره دسته	همخوان‌ها	شماره دسته	همخوان‌ها
۱	p , t , k	٤	s , š , č , f
۲	b , d , g	٥	z , j , ž , q , v
۳	m , n , l , r , y	٦	h , x , ?

یاد آوری- برای نمایش واج‌ها از الفبای واجی مرجع [۵۰] استفاده شده است.

در لایهٔ خروجی نیز اطلاعات نوایی هجای جاری بدین‌ترتیب تولید می‌شوند:

الف) چهار پارامتر (ضرایب مراتب صفر تا سوم بسط چند جمله‌ای لژاندر گسسته) مبین الگوی گام [51].

ب) چهار پارامتر (ضرایب مراتب صفر تا سوم بسط چند جمله‌ای لژاندر گسسته) مبین الگوی انرژی (در مقیاس لگاریتمی).

پ) یک پارامتر مبین میزان "وقفه قبل از هجا" (برحسب میلی‌ثانیه).

ت) سه پارامتر مبین میزان دیرش قطعات "هجا" و "واکدار هجا" و نیز "نقطهٔ آغاز قطعه واکدار" (برحسب میلی‌ثانیه).

تعداد نرون‌ها در لایه‌های مخفی اول و دوم نیز به صورت تجربی به ترتیب ۳۳ و ۳۰ در نظر گرفته شده‌اند.

برای کاهش پیچیدگی سیستم نیز نرون‌ها در لایهٔ مخفی دوم و خروجی به سه گروه تفکیک شده‌اند.

بدین‌ترتیب ساختار تفصیلی RNN مولد نوا به صورت شکل (۶) قابل نمایش است.

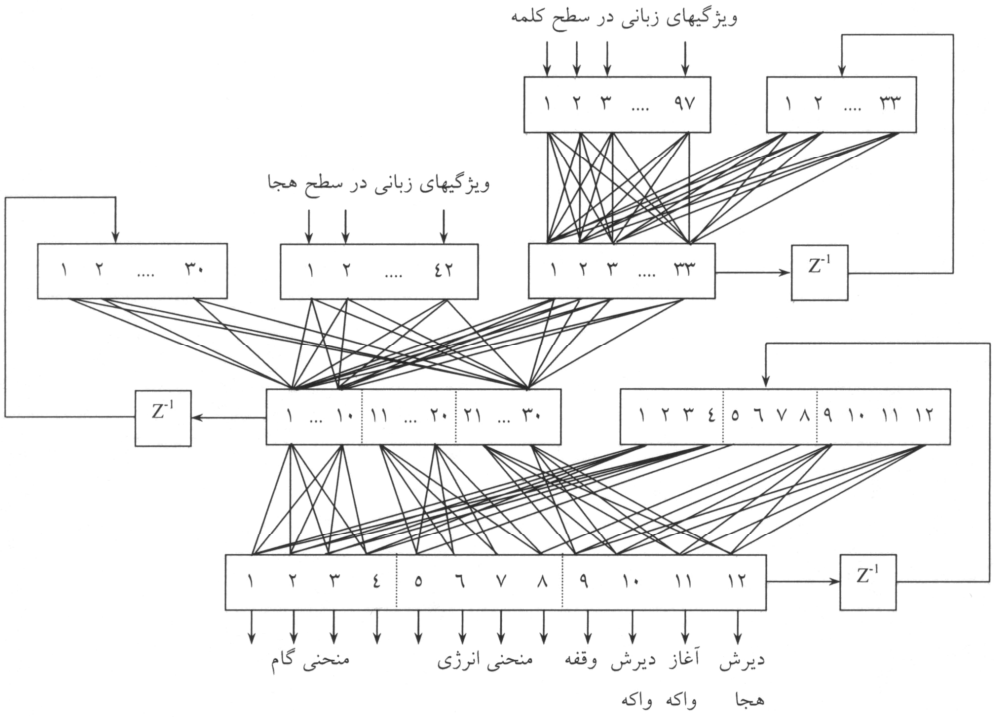
اطلاعات آموزشی RNN از ۴۰۰ جمله "مثبت" و "منفی" (از انواع "خبری"، "پرسشی"، "امر" / "نهی" و "تعجبی") "کوتاه" و "بلند" که توسط یک گوینده مرد با میانگین سرعت چهار هجا بر ثانیه ادا شده‌اند، بدست آمده است. به عنوان ابزار شبیه‌سازی سیستم نیز از Toolbox شبکه عصبی نرم‌افزار MATLAB بهره گرفته شده است.

نرخ‌های یادگیری برای دو نوع ضرایب وزن متصل به نرون‌های مخفی و خروجی در ابتدا ۰/۰۱ و ۰/۰۰۱ در نظر گرفته شده که طی ۲۰۰ دور آموزشی (Epoch) به صورت خطی کاهش می‌یابند تا به صفر برسند. فرآیند آموزش نیز پس از ۵۳ دور همگرا شده است. شکل (۷) عملکرد سیستم مولد نوا را در پیش‌بینی الگوی گام و انرژی بهنگام ادای یک جمله (از ۱۰۰ جمله بکار رفته برای آزمون) نشان می‌دهد.

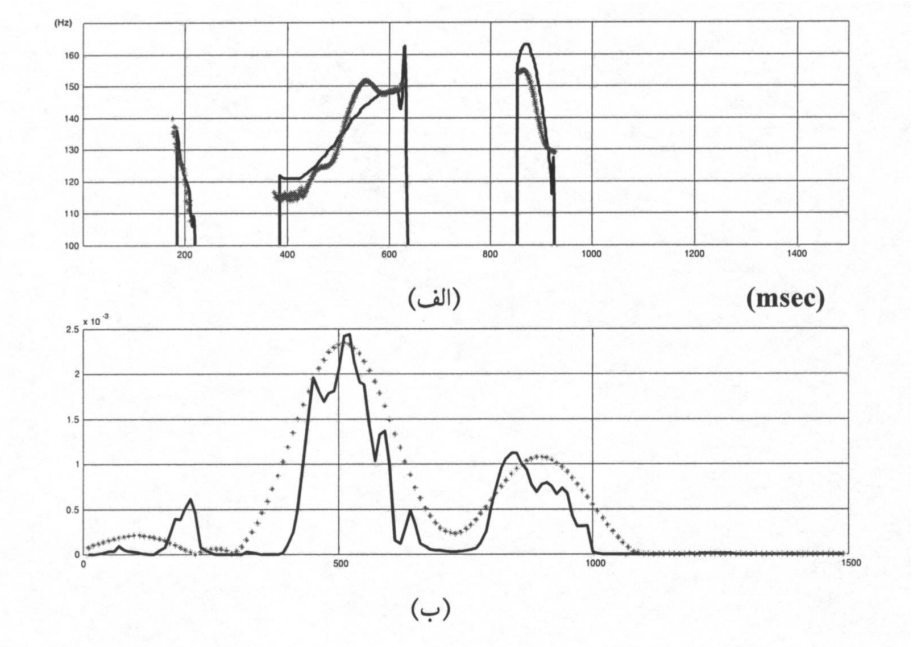
در جدول (۴)، RMSE (Root Mean Square Error) مربوط به چهار پارامتر نوا برای ۱۰۰ جملهٔ آزمون فهرست شده است.

جدول ٤ RMSE چهار پارامتر نوای سنتز شده

نام پارامتر	RMSE
الگوی گام	۰/۹٦ ms/Frame
الگوی انرژی	۲/۲٥ dB
دیرش واکه	۳۸/۱ ms
دیرش وقفه	٤۷/۳ ms



شکل ٦ ساختار تفصیلی RNN مولد نوای گفتار فارسی



شکل ٧ عملکرد سیستم مولد نوای گفتار ، الف) الگوی فرکانس گام، ب) الگوی انرژی
(خط پر: الگوی واقعی، *: الگوی سنتز شده)

سنتز گفتار

دیدیم که برای سنتز گفتار، روش‌های مبتنی بر قاعده [52-54] و نیز اتصال قطعات گفتار [23,55-61] و همچنین آمیختار (به عنوان نمونه سیستمی که با

بکارگیری مدل سری / موازی Klatt، از روش اتصال قطعات نیز در مرز بین قطعات واکدار بهره می‌جوید [62]) ارائه شده‌اند.

در سیستم TTS حاضر از روش HNM (نوع HNM1 [63]) برای سنتز به روش اتصال قطعات استفاده شده است [65-64]. در این رویکرد الگوریتم‌های زیر پیاده‌سازی شده‌اند:

الف) تخمین گام و حداکثر فرکانس واکداری MVF (Maximum Voiced Frequency) [66]

ب) تعیین دامنه و فاز هارمونیک‌ها [65]

پ) تصحیح فاز هر فریم آنالیز (الگوریتم مرکز ثقل (Center of gravity) [67]

ت) هموارسازی(Smoothing) پارامترهای HNM حول نقاط اتصال قطعات گفتار [68]

ث) اصلاح نوا (فرکانس گام و دیرش) و تعیین لحظات زمانی سنتز مربوط [70-69]

ج) تخمین مقادیر دامنه و فاز هارمونیک‌ها با فرکانس گام جدید [71-63,64]

چ) هموارسازی ناپیوستگی‌های حاصل از عدم انطباق سه عامل فاز، گام و پوش طیفی در محل اتصال دو واکه (در تولید هجاهای CVC و CVCC)

ح) سنتز قطعات گفتاری [64] (البته در این سیستم برای آنالیز و سنتز قسمت‌های نویزی، یک طول گام مجازی به میزان دو برابر نقاط زمانی آنالیز در نظر گرفته شده و بدین‌ترتیب آنالیز و سنتز نواحی نویزی نیز همچون نواحی هارمونیکی صورت گرفته است [72]).

پایگاه داده واحدهای گفتاری، با توجه به ساختار هجایی زبان فارسی و تعداد واکه‌ها و همخوان‌های این زبان (به ترتیب ۶ واکه و ۲۳همخوان) از ۱۳۸ دو واجی از نوع CV، ۱۳۸ دو واجی از نوع VC و ۲۳ همخوان (در مجموع ۲۹۹ قطعه به عنوان یک مجموعه) تشکیل شده است.

در این بخش از یک منحنی پیوسته تکه‌ای خطی (Piecewise linear) برای تعیین تغییرات مقادیر فرکانس گام برحسب هرتز استفاده شده است(مشابه سنتز کننده(Multi-Band Resynthesis OverLap Add) MBROLA [73]). اطلاعات تغییر مقادیر فرکانس گام و دیرش در سطح هجا نیز به صورت زیر به سنتز کننده اعمال می‌شوند:

[100 ... 0]: نشانه‌های زمانی گام (برحسب درصدی از طول ناحیه واکدار)

[f_1 ... f_N]: مقادیرگام (برحسب هرتز) (برحسب میلی ثانیه)

دیرش هر یک از واج‌های هجا]: دیرش واج‌های هجا

گفتنی است که در مورد قطعات بی‌واک دیرش را تغییر نداده و از عدد صفر به نشانه عدم تغییر دیرش استفاده شده است. در این طرح از یک الگوریتم بازگشتی مشابه آنچه در طراحی فیلترهای گسسته بهینه بکار گرفته می‌شود [74]، برای تعیین مکان و فرکانس لحظات زمانی سنتز استفاده شده است. حدس اولیه در این الگوریتم، تعداد نقاط سنتز است که با توجه به مقدار میانگین فرکانس گام در طول هجا و همچنین دیرش هجای مورد نظر بدست می‌آید. شکل (۸) به عنوان نمونه هجای /čci/ را قبل از اصلاح نوای گفتار و پس از آن (به همراه منحنی گام مربوط) نشان می‌دهد.

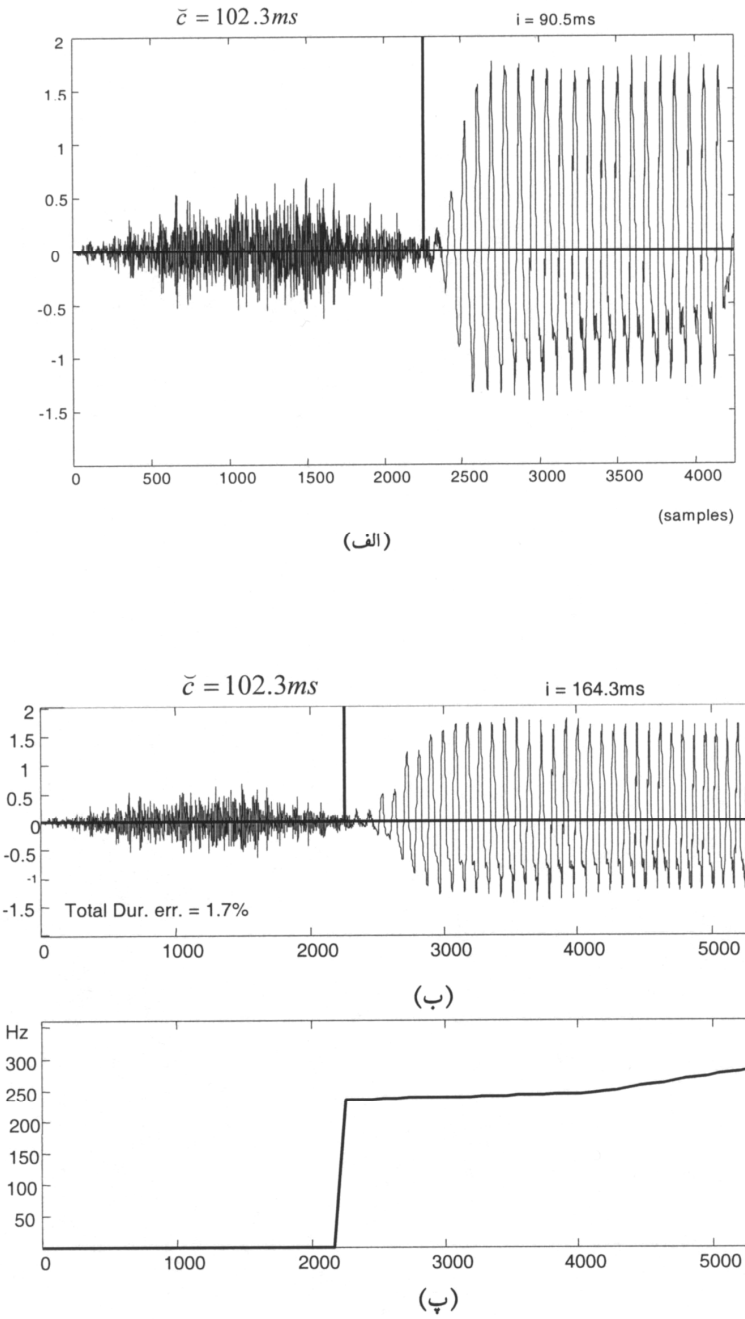
امکان تغییر مقادیر بلندی صدا (Loudness) نیز در سیستم یاد شده در نظر گرفته شده است (علی‌رغم اهمیت کمتر آن در مقایسه با مقادیر گام و دیرش). این کار با تنظیم مقادیر انرژی فریم‌ها و بکارگیری یک منحنی پیوسته تکه‌ای خطی از ضرایب تصحیح مقادیر انرژی در سطح هجا صورت می‌پذیرد. نحوه اعمال اطلاعات تغییر مقدار انرژی نیز به صورت زیر است:

[100 ... 0]: نشانه‌های زمانی انرژی (برحسب درصدی از طول هجا)

[k_1 ... k_N]: ضریب انرژی (ضرایب ثابت تضعیف $0 \le k_i \le 1$)

ارزیابی عملکرد سیستم

ارزیابی عملکرد سیستم طراحی شده با توجه به استاندارد P.85 از ITU-T انجام شده است [75].



شکل ۸ نتایج تغییرات نوا بر هجای / $\breve{c}i$ / ، الف) قبل از تغییر نوا، ب) پس از تغییر نوا، پ) منحنی گام اصلاحی

ملاک‌های ارزیابی بکار رفته با مقیاس MOS (Mean Opinion Score) عبارتند از: نظر کلی (Overall impression) (درباره کیفیت صدا)، تلاش شنیداری (Listening effort) (برای درک پیام)، میزان درک (Comprehension) (مشکل در فهم کلمات)، نحوه ادا (Articulation) (میزان تمایز بین صداها)، کیفیت تلفظ (Pronunciation)، خوشایندی (Pleasantness) صدا. از ۲۰ شنونده (با سنین ۱۵ تا ۴۵ سال) برای اظهارنظر در مورد ۲۴ جمله (با حداقل دیرش ۱۰ ثانیه استفاده شده و به دلیل تعدد ملاک‌ها، هر جمله سنتز شده دوبار

جدول ۵ نتایج ارزیابی عملکرد سیستم FTTS و ٦ سیستم نمونه جهانی

ملاک ارزیابی	نام سیستم						
	EL	LT	AK	RS	SS	ATT	FTTS
نظر کلی	۱/۷۳	۲/۱۹	۲/۲۸	۳/۱۳	۳/٦۰	۳/۷٤	۳/٦۸
تلاش شنیداری	۲/۲۸	۲/۸۱	۲/۸۳	۳/٤۳	۳/۷۰	۳/۷۲	۳/٦٥
میزان درک	۲/٤۷	۲/۹۷	۳/۰۲	۳/٦۳	۳/۸۸	۳/۹۱	۳/۹۷
نحوهٔ ادا	۲/۲۱	۲/٦٥	۲/٦۷	۳/۳٤	۳/٦۹	۳/۸۱	۳/۷٤
کیفیت تلفظ	۲/۰٤	۲/٦۰	۲/٦۰	۳/۱۷	۳/٤۹	۳/۳۰	۳/۳۷
خوشایندی صدا	۲/۰۱	۲/۱۰	۲/٤۹	۳/۰٥	۳/٦۰	۳/٦٥	۳/۱۲
برآورد کلی (میانگین ملاک‌های فوق)	۲/۱۲	۲/٥٥	۲/٦٥	۳/۲۹	۳/٦٦	۳/٦۹	۳/٥۹

پخش شده تا در بار اول در مورد سه ملاک اول و در بار دوم در مورد سه ملاک دوم اظهارنظر شود.

نتایج مزبور در مورد سیستم حاضر که آن را FTTS (Farsi TTS) نامیده‌ایم و نیز شش سیستم TTS مدرن (Lucent ،EL (Elan Informatique ،LT echnology) ، AK(Aculab) ،RS (Real Speak) ،SS (Speechworks Speechify) و (AT&T Next ATT Generation) که در شرایط نسبتاً مشابهی مورد ارزیابی قرار گرفته‌اند [76]، در جدول (۵) آورده شده است.

البته به خاطر تفاوت بین زبان‌ها در FTTS و شش سیستم یاد شده ممکن است به مقایسه فوق اشکال وارد باشد، ولی می‌توان این نتیجه را گرفت که با توجه به نبود یک سیستم TTS جامع برای زبان فارسی تا زمان نگارش مقالهٔ حاضر [77]، عملکرد FTTS قابل مقایسه با سیستم مشابه دیگری در زبان فارسی نبوده، اما عملکرد آن در زمرهٔ سیستم‌های موفق امروزی در این زمینه می‌باشد.

نتیجه‌گیری

یک سیستم تبدیل متن به گفتار طبیعی برای زبان فارسی در قالب بلوک‌های کاری زیر طراحی و پیاده‌سازی شد:

الف) پردازشگر زبان طبیعی (با دو خروجی "رشته‌های واجی" و" اطلاعات نحوی کلمات متن")

ب) مولد نوای گفتار (برای پیش‌بینی الگوی فرکانس گام، الگوی انرژی، اطلاعات دیرش و وقفه در سطح هجا و اجزای آن)

پ) سنتزکننده گفتار (به روش HNM و البته با اصلاحاتی در زمینه تغییر نوا)

برای تهیه سریع و خودکار اطلاعات لازم برای آموزش شبکه عصبی مولد نوای گفتار نیز روش‌های ابتکاری برای تقطیع و نامگذاری قطعات گفتاری با دقت بالا ارائه شد. علمکرد سیستم بر اساس استاندارد P.85 از ITU-T ارزیابی گردید. میانگین MOS در شش ملاک بررسی شده، ۳/۵۹ می‌باشد که در ردهٔ سیستم های مدرن TTS برای زبان انگلیسی است. تلاش در زمینه ایجاد تنوع در قطعات گفتاری در پایگاه دادهٔ سنتز، تحلیل معنایی در بخش NLP (به منظور رفع ابهام تمامی کلمات فارسی) و نیز بی‌درنگ نمودن عملکرد سیستم ادامه دارد.

مراجع

1. Sheikhan M., et al., "Continuous speech recognition and syntactic processing in Iranian Farsi Language", *Int. J. Speech Technology, Kluwer Academic Publishers*, pp. 135-141, (1997).

2. Sagisaka Y., "Speech synthesis from text", *IEEE Commun. Mag*, pp. 35-41, (1990).

3. Sheikhan M., et al., "Using symbolic and connectionist approaches to automate editing Persian sentences syntacticly", *Proc. Int. Conf. on Intelligent and Cognitive Systems*, pp. 250-253, (1996).

٤. وحیدیان کامیار ت.، *نوای گفتار در فارسی*، انتشارات دانشگاه جندی شاپور، ١٣٥٧.

5. Allen J., "*Overview of TTS systems*", in Advances in Speech Signal Processing, Marcel Dekker, (1991).

6. Thorensen N., "Sentence intonation in textual context-supplementary data", *J. Acoust. Soc. Am., 80 (4)*, pp. 1041-1047, (1986).

7. Sagisaka Y., "On the prediction of global F0 shape for Japanese TTS", *Proc. ICASSP' 90*, pp. 325-328, (1990).

8. Buhmann J., et al., "Intonation modeling for the synthesis of structured documents", *Proc. ICSLP' 2002*, pp. 2089-2092, (2002).

9. Riedi M., "A neural-network-based model of segmental duration for speech synthesis", *Proc. Eurospeech' 95*, pp. 599-602, (1995).

10. Yiqing Z., "Syllable duration and its functions in standard Chinese discourse", *Proc. ICSLP' 2000*, Paper No. 1097, (2000).

11. Smith C.L., "Modeling durational variability in reading aloud a connected text*", Proc. ICSLP' 2002*, pp. 1769-1772, (2002).

12. Sagisaka Y. and Sato H., "*Accentuation rules in Japanese TTS conversion*", Rev. Elect. Commun. Lab., 32 (2), pp. 188-199, (1984).

13. Low P. H. and Vaseghi S., "Application of microprosody models in TTS synthesis", *Proc. ICSLP' 2002,* pp. 2413-2416, (2002).

14. Hifny Y. and Rashwan M., "Duration modeling for Arabic TTS synthesis", Proc. *ICSLP' 2002*, pp. 1773-1776, (2002).

15. Mixdorf H. and Fujisaki H., "A scheme for a model-based synthesis by rule of F0 contours of German utterances", *Proc. Eurospeech' 95*, pp. 1823-1826, (1995).

16. Lopez-Gonzale E. and Hernandez-Gomez L.A., "Automatic data-driven prosodic modeling for TTS", *Proc. Eurospeech' 95*, pp. 585-588, (1995).

17. Yamashita Y. and Mizoguchi R., "Modeling the contextual effects on prosody in dialog*", Proc. Eurospeech' 95*, pp. 1329-1332, (1995).

18. Ostendorf M. and Veilleux N., "A hierarchical stochastic model for automatic prediction of prosodic boundary location", *Computat. Linguist*., 20, pp. 27-54, (1994).

19. Chen S. H., et al., "An RNN-based prosodic information synthesizer for Mandarin TTS", *IEEE Trans. Speech Audio Processing, 6 (3),* pp. 226-239, (1998).

20. Sheikhzadeh H., et al., "Farsi language prosodic structure, research and implementation using a speech synthesizer", *Proc. Eurospeech' 99, Vol. 4,* pp. 1647-1650, (1999).

21. Kinoshita K., et al., "Duration and F0 as perceptual cues to Japanese vowel quantity", *Proc. ICSLP' 2002,* pp. 757-760, (2002).

22. Breen A., "Speech synthesis models: a review", *Elect. Commun. Engng. J.*, pp. 19-31, Feb. (1992).

23. Syrdal A. K., et al., "Corpus-based techniques in the AT&T NEXTGEN synthesis system", *Proc. ICSLP' 2000*, Paper No. 3601, (2000).

٢٤. نصیرزاده م. و صیادیان ا.، "تجربه‌ای در مدل‌سازی زبان فارسی برای یک سیستم تبدیل متن به گفتار"، مجموعه مقالات دومین کنفرانس سالانه انجمن کامپیوتر ایران، صفحات ١١١-١٠٥، (١٣٧٥).

25. Childers D. G., Speech Processing and Synthesis Toolboxes, Wiley, (2000).

٢٦. فرخی ع. و همکاران، "تقطیع خودکار سیگنال گفتار در سطوح مختلف جمله به منظور استخراج اطلاعات نوا"، مجموعه مقالات دهمین کنفرانس مهندسی برق ایران (گرایش کامپیوتر)، صفحات ٤٣١-٤٢٤، (١٣٨١).

27. Sethy A. and Narayanan S., "Refined speech segmentation for concatenative speech synthesis", *Proc. ICSLP' 2002*, pp. 149-152, (2002).

28. Klatt D. H., "Review of TTS conversion for English", *J. Acoust. Soc. Am., 82 (3),* pp. 737-793, (1987).

29. Yiourgalis N., et al., "Some important cues on improving the quality of a TTS system", *Proc. ICSST' 92*, pp. 528-533, (1992).

30. O'Malley M. H., et al., "An analysis of strategy for finding prosodic cues in text", *Proc. Eurospeech' 91,* pp. 1165-1168, (1991).

31. Horvei B., et al., "Analyzing prosody by means of a double tree structure", *Proc. Eurospeech' 93*, pp. 1987-1990, (1993).

32. Traber C., "Syntactic processing and prosody control in the SVOX TTS system for German*", Proc. Eurospeech' 93*, pp. 2099-2102, (1993).

33. Baily G., et al., "Integration and rhythmic and syntactic constraints in a model of generation of French prosody", *Speech Commun., 8*, pp. 137-146, (1989).

34. Frid J., "Prediction of intonation patterns of accented words in a corpus of read Swedish news through pitch contour stylization", *Proc. Eurospeech' 2001*, pp. 915-918, (2001).

35. Mittrapiyanuruk P., et al., "Improving naturalness of Thai TTS synthesis by prosodic rule", *Proc. ICSLP' 2000*, Paper No. 190, (2000).

36. Kaiki N., et al., "Statistical modeling of segmental duration and power control for Japanese", *Proc. Eurospeech' 91*, pp. 625-628, (1991).

37. Fukuda T., et al., "A study of pitch pattern generation using HMM-based statistical information", *Proc. ICSLP' 94*, pp. 723-726, (1994).

38. Fujio S., et al., "Stochastic modeling of pause insertion using context-free grammar", *Proc. ICASSP' 95*, pp. 604-607, (1995).

39. Saito T. and Sakamoto M., "Generating F0 contours by statistical manipulation of natural F0 shapes", *Proc. Eurospeech' 2001*, pp. 1171-1174, (2001).

40. Sun X. and Applebaum T. H., "Intonation phrase break prediction using decision tree and N-gram *model", Proc. Eurospeech' 2001*, pp. 537-540, (2001).

41. Yamashita Y. and Ishida T., "Stochastic F0 contour model based on the clustering of F0 shapes of a syntactic unit", *Proc. Eurospeech' 2001*, pp. 533-536, (2001).

42. Donovan R. E., "A component by component listening test anlysis of the IBM trainable speech synthesis system", *Proc. Eurospeech' 2001*, pp. 329-332, (2001).

43. Kim W., et al., "Model-based stress decision method*", Proc. Eurospeech' 2001*, pp. 107-110, 2001.

44. Scordilis M. S., et al., "Neural network-based generation of fundamental frequency contours", Proc. ICASSP' 89, pp. 219-222, (1989).

45. Taylor P., "Using neural networks to locate pitch accents", *Proc. Eurospeech' 95*, pp. 1345-1348, (1995).

46.Call D., et al., "Neural processes underlying perceptual learning of difficult second language phonetic contrast", *Proc. Eurospeech 2001*, pp. 145-148, (2001).

47.Mueller A. F. and Hoffmann R., "Accent label prediction by time delay neural network using gating clusters", *Proc. Eurospeech' 2001*, pp. 549-552, (2001).

48.Caglayan E., et al., "Natural F0-contours with a new neural-network-hybrid approach", *Proc. ICSLP' 2000,* Paper No. 1172, (2000).

49.Sheikhan M., "RNN-based prosodic information synthesizer for Farsi TTS", Second Irano-Armenian Workshop on Neural Networks, Dec. (1999).

۵۰. ثمره ی.، آواشناسی زبان فارسی، مرکز نشر دانشگاهی، (۱۳۷٤).

51. Chen S. H. and Wang Y. R., "Vector quantization of pitch information in Mandarin speech", *IEEE Trans. Commun., 38,* pp. 1317-1320, (1990).

52. Allen J., et al., From Text to Speech: The MITalk System, Cambridge University Press, (1987).

53. Carlson R., et al., "A multilanguage TTS system", *Proc. ICASSP' 82 Vol. 3*, pp 1604-1607, (1982).

54. Fujisaki H., et al., "A system for synthesizing Japanese speech from orthographic text", *Proc. ICASSP' 90,* Vol. 1, pp. 617-620, (1990).

55. Huang X., et al., "Whistler: A trainable TTS system", *Proc. ICSLP' 96*, Vol. 4, pp 2387-2390, 1996.

56. Kawai H., et al., "Development of a TTS for Japanese based on waveform splicing", *Proc. ICASSP' 94,* Vol. 1, pp. 569-572, (1994).

57. Yamada M., et al., "PURETALK: A high quality Japanese TTS system", *Proc. ICSLP' 2000*, Paper No. 875, (2000).

58. Moulines E., et al., "A real-time French TTS system generating high quality synthetic speech", *Proc. ICASSP' 90, Vol. 1*, pp. 309-312, (1990).

59. Matousek J. and Psutka J., "ARTIC: A new Czech TTS system using statistical approach to speech segment database construction", *Proc. ICSLP' 2000*, Paper No. 444, (2000).

60. Christogiannis C., et al., "Design and implementation of a Greek TTS system based on concatenative synthesis", *Proc. ICSLP' 2000*, Paper No. 404, (2000).

61. Bulut M., et al., "Expressive speech synthesis using a concatenative synthesizer", *Proc. ICSLP' 2002*, pp. 1265-1268, (2002).

62. Yiourgalis N. and Kokkinakis G., "TTS for Greek", *Proc. ICASSP' 91,* Vol. 1, pp. 525-528, (1991).

63. Stylianou Y., "Concatenative speech synthesis using a harmonic plus noise model", 3rd ESCA / COCOSDA Workshop on Speech Synthesis, Nov. (1998).

64. Stylianou Y., et al., "High quality speech modification based on a harmonic + noise model", *Proc. Eurospeech' 95*, pp. 451-454, (1995).

65. O'Brien D. and Monaghan A., "Concatenative synthesis based on a harmonic model", *IEEE Trans. Speech Audio Processing, 9 (1),* pp. 11-20, (2001).

66. Stylianou Y., "A pitch and maximum voiced frequency estimation technique adapted to harmonic models of speech", *IEEE Nordic Signal Processing Symposium*, (1996).

67. Stylianou Y., "Removing phase mismatches in concatenative speech synthesis", Third ESCA Speech Synthesis Workshop, Nov. (1998).

68. Syrdal A., et al., "TD-PSOLA versus harmonic plus noise model in diphone based speech synthesis", *Proc. ICASSP,* (1998).

69. Moulines E. and Charpentier F., "Pitch synchronous waveform processing techniques for TTS synthesis using diphones", *Speech Commun*., 9, pp. 453-467, (1990).

70. Quatieri T. F. and McAulay R. J., "Shape invariant time-scale and pitch modification of speech", *IEEE Trans. Signal Processing, 40*, pp. 497-510, (1992).

71. Dutiot T., An Introduction to Text-To-Speech Synthesis, Kluwer Academic Publishers, (1996).

72. O'Brien D. and Monaghan A., "Shape invariant time-scale modification of speech using a harmonic model", *Proc. ICASSP' 99,* pp. 381-384, (1999).

73. *http://tcts.fpms.ac.be/synthesis/mbrola.htm*.

74. Oppenheim A. V. and Schafer R. W., Discrete-Time Signal Processing, Prentice-Hall, (1989).

75. ITU-T Recommendation P.85, "A method for subjective performance assessment of the quality of speech output devices*", International Telecommunications Union publication*, (1994).

76. Alvarez Y. V. and Huckvale M., "The reliability of the ITU-T P.85 standard for the evaluation of TTS systems", *Proc. ICSLP' 2002*, pp. 329-332, (2002).

77. Mohammadi M. and Sheikhan M., "TTS in broadcasting", *Proc. Int. Conf. BroadcastAsia 2000*, Singapore, (2000).