تولید خودکار نوای گفتار به‌کمک مدل آمیختار عصبی- آماری با امکان انتخاب واحد در سنتز

منصور شیخان

استادیار گروه مخابرات، دانشکدهٔ فني و مهندسي، واحد تهران جنوب دانشگاه آزاد اسلامي، تهران

چکیده

در این مقاله با هدف ایجاد بهبود در عملکرد اولین ویرایش از سیستم تبدیل متن به گفتار طبیعی ارائه شده برای زبان فارسی، که در آن از یک شبکهٔ عصبی بازگشتی برای تولید همزمان عوامل نوای گفتار (الگوی فرکانس گام، دیرش، انرژی و درنگ) و نیز سنتزکنندهٔ "مدل هارمونیک + نویز" با دادگان تک‌واحدی از دوواجی‌ها، برای تولید گفتار استفاده شده بود، چگونگی به‌کارگیری یک مدل آمیختار عصبی- آماری برای تولید نوا با ایجاد امکان انتخاب واحد در بخش سنتز گفتار بررسی شده است.

در این راستا، در مدل آمیختار توسعه‌یافته برای تولید نوا، شبکهٔ عصبی بازگشتی نقش پارامتری ساختن عوامل نوا (البته با تغییراتی در ساختار، نسبت به ویرایش اول سیستم) را داشته و به دنبال آن درخت‌های تصمیم‌گیری نوایی شکل گرفته و توصیف توزیع مقادیر نوا نیز به‌وسیلهٔ مدل مخلوط گوسی انجام شده است. بدین ترتیب امکان انتخاب قطعات گفتاری بهینه در یک سیستم سنتز مبتنی بر انتخاب واحدهای هجایی و اتصال قطعات مناسب فراهم آمد. در ارزیابی نهایی عملکرد سیستم نیز با به‌کارگیري ملاك‌هاي ارائه شده در استاندارد ITU-T P.85، MOS معادل ۳/۶ محاسبه شد.

واژه‌هاي کلیدي: نوای گفتار؛ الگوی گام؛ دیرش؛ شبکهٔ عصبی بازگشتي؛ درخت تصمیم‌گیری؛ مدل مخلوط گوسی؛ سنتزگفتار؛ انتخاب قطعهٔ گفتاری؛ هجا

---

[*] عهده‌دار مکاتبات

نشاني: تهران، خیابان ملاصدرا، خیابان شیراز جنوبی، کوچهٔ سرو، شمارهٔ۷، مرکز تحصیلات تکمیلی دانشگاه آزاد اسلامی- واحد تهران جنوب

تلفن: ۹۰-۸۸۰۵۹۶۸۷، ۰۹۱۲۱۱۶۳۱۳۲، **دورنگار:** ۸۸۰۵۹۶۹۰، **پیام نگار:** msheikhn@azad.ac.ir

۱- مقدمه

در دو دهة اخیر تلاش‌هایی برای ایجاد ارتباط طبیعی بین انسان و رایانه و توسعة سیستم‌هایی برای کمک‌رسانی به افراد ناتوان جسمی، با قابلیت به‌کارگیري در سایر حوزه‌های کاربردی، در پژوهش‌های داخل کشور مشاهده می‌شود. در راستای برقراری این ارتباط می‌توان به سیستم‌های بازشناسی گفتار پیوسته (CSR)[۱] برای کمک به ناشنوایان و سیستم‌های تبدیل متن به گفتار (TTS)[۲] برای کمک به نابینایان اشاره کرد [۱-۹].

در این مقاله، هدف معرفی رویکرد اصلاحی برای بهبود عملکرد نسخة اول سیستم TTS ارائه شده به‌وسیلة مؤلف و همکاران در مرجع [۶] برای زبان فارسی است.

در سیستم TTS توسعه‌یافته در مرجع [۶]، یک پردازشگر زبان طبیعی (NLP)[۳] ، عملیات تعیین مشخصة دستوری کلمات (POS)[۴] [۱۰] و تبدیل متن به واج (همراه با رفع مشکل کسرة اضافه برای اغلب ترکیب‌های اضافی در زبان فارسی) را انجام می‌داد. برای تولید نوا نیز از یک شبکة عصبی بازگشتی (RNN)[۵] برای تخمین همزمان پارامترهای نوای گفتار استفاده مي‌شد. سنتز گفتار نیز با دادگان تک‌واحدی (SU)[۶] از دوواجی‌ها[۷] به روش "مدل هارمونیک + نویز" (HNM)[۸] صورت می‌پذیرفت.

به منظور بهبود عملکرد TTS مذکور با واگذاری نقش پارامتری ساختن عوامل نوا[۹] (یعنی الگوی گام[۱۰]، دیرش[۱۱]، انرژی و درنگ[۱۲]) به RNN (البته با تغییراتی در ساختار، نسبت به RNN ارائه شده در مرجع [۶]) و تشکیل درخت‌های تصمیم‌گیری[۱۳] نوایی و توصیف توزیع مقادیر نوا به‌وسیلة مدل مخلوط گوسی (GMM)[۱۴]، مدل آمیختاری برای نوا به‌کار گرفته شده است که امکان انتخاب قطعات گفتاری را در دنبالة سنتز فراهم می‌آورد. بدین‌ترتیب در مقایسه با سیستم TTS ارائه شده در مرجع [۶]، در پژوهش گزارش شده در مقالة حاضر، دو تغییر عمده اعمال شد:

الف- به‌کارگیری مدل تلفیقی عصبی-آماری برای تولید نوا؛

ب- امکان سنتز مبتنی بر انتخاب قطعات هجایی گفتار، به جای اتصال قطعات دوواجی از دادگان تک‌واحدی.

۲- روش‌ها

۲-۱- مدل آماری نوا با هدف انتخاب واحد بهینة گفتاری

مدل‌های نوا وظیفة تولید اطلاعات الگوی فرکانس گام [۱۱-۱۴]، دیرش [۱۵-۱۹]، انرژی [۲۰،۲۱] و درنگ [۲۲] را دارند. تاکنون از سیستم‌های مبتنی بر قاعده[۱۵] [۲۳-۲۵] و مبتنی بر آموزش[۱۶] (مدل‌های آماری [۲۶-۳۲]، شبکه‌های عصبی [۳۳-۳۷]، الگوریتم ژنتیک [۳۸] و مدل‌های آمیختار[۱۷] [۳۹،۴۰]) برای تولید اطلاعات نوا بهره گرفته شده است.

در این پژوهش از مدل‌های آماری (که در آن مقادیر پیش‌بینی شده برای عوامل نوا، با توزیع احتمالاتی ارائه خواهند شد) با هدف انتخاب واحد بهینة گفتاری در سیستم TTS فارسی استفاده شده است. در این راستا توضیحات مربوط به مدل‌های نوایی "هدف"[۱۸]، "گذار"[۱۹] و نیز چگونگی انتخاب واحد بهینة گفتاری به همراه روابط پایة آماری مربوط در ادامة این بخش آورده شده است.

۲-۱-۱- مدل نوایی هدف

در این مورد فرض کنید که $S_i (1 \le i \le N)$ دنبالة واحد سنتز، $CTX_i (1 \le i \le N)$ توصیف بافت[۲۰] واحد S_i، $C(S_i, j)(1 \le j \le M_i)$ هم jاُمین واحد کاندید S_i (که M_i تعداد کل کاندیدهای واحد S_i است)، $P_k(C(S_i, j))\ (1 \le k \le L)$ نیز k اُمین بردار مبین اهداف نوایی $C(S_i, j)$، و $Z_k(CTX_i)(1 \le k \le L)$ توزیع در خوشة با بافت یکسان برای k اُمین بردار هدف نوا باشد.

بر این اساس، احتمال هدف اینکه $C(S_i, j)$ به‌عنوان نمونة S_i انتخاب شود، چنین تعریف می‌شود [۴۱-۴۳]:

$$P(C(S_i, j) \mid S_i, CTX_i) = \prod_{k=1}^{L} P(P_k(C(S_i, j)) \mid Z_k(CTX_i)) \qquad (۱)$$

نکتة قابل توجه این است که در این مقاله یک فرایند دو مرحله‌ای برای مدل‌سازی سمت چپ رابطة (۱) به‌کار گرفته شده است. در فاز آموزش، ویژگی‌های بافت به‌وسیلة

---

[1] **C**ontinuous **S**peech **R**ecognition
[2] **T**ext **T**o **S**peech
[3] **N**atural **L**anguage **P**rocessor
[4] **P**art **O**f **S**peech
[5] **R**ecurrent **N**eural **N**etwork
[6] **S**ingle **U**nit
[7] Diphones
[8] **H**armonic + **N**oise **M**odel
[9] Prosody
[10] Pitch contour
[11] Duration
[12] Pause
[13] Decision trees
[14] **G**aussian **M**ixture **M**odel
[15] Rule-based
[16] Data-driven
[17] Hybrid
[18] Target
[19] Transition
[20] Context

درخت تصمیم‌گیری و بر اساس توزیع ویژگی‌های نوایی ارائه شده به‌وسیلة RNN، خوشه‌بندی می‌شوند. به‌بیان دیگر ویژگی‌های نوایی مشابه در یک گره برگ دسته‌بندی می‌شوند. سپس برای هر برگ، مدل مخلوط گوسی به‌منظور مدل‌سازی توزیع احتمالاتی ویژگی‌های نوایی، آموزش داده می‌شود.

۲-۱-۲- مدل نوایی گذار

از نظر درک گفتار، واریانس‌های پارامترهای نوایی دو دنبالة متوالی، نقش مهمی را در طبیعی بودن گفتار بازی می‌کنند و بر همین اساس مدل نوایی گذار را به‌کار می‌گیرند [۴۱].

در این مورد فرض کنید که $C(S_{i-1}, j)(1 \le j \le M_{i-1})$ واحد کاندید j اُم برای S_{i-1}، $C(S_i, \ell)(1 \le \ell \le M_i)$ هم ℓ اُمین واحد کاندید برای S_i، $Q_k(C(S_{i-1}, j), C(S_i, \ell))(1 \le k \le T)$ نیز k اُمین بردار مبین گذارهای نوایی بین $C(S_{i-1}, j)$ و $C(S_i, j)$، و $T_k(CTX_i)(1 \le k \le T)$ توزیع در خوشة با بافت یکسان برای k اُمین بردار گذار نوا باشد.

بر این اساس، احتمال اینکه گذار $C(S_{i-1}, j)$ و $C(S_i, \ell)$ به‌عنوان نمونه‌های S_{i-1} و S_i انتخاب شوند، به‌صورت رابطة (۲) تعریف می‌شود [۴۱-۴۳]:

$$P(C(S_{i-1}, j), C(S_i, \ell) | S_{i-1}, S_i, CTX_i) = \prod_{k=1}^{T} P(Q_k(C(S_{i-1}, j), C(S_i, \ell)) | T_k(CTX_i)) \qquad (۲)$$

۳-۱-۲- چارچوب انتخاب واحد بهینة گفتاری

برای هر کاندید از واحد سنتز، هر چه احتمال هدف بالاتر باشد، آن کاندید بهتر است. به همین ترتیب برای هر ترکیبی از دو کاندید برای دو واحد سنتز متوالی، هر چه احتمال گذار بالاتر باشد، کاندیدهای مذکور بهتر خواهند بود. بر این اساس مسالة انتخاب واحد گفتاری، معادل یافتن مسیر بهینه برای دنبالة سنتز خواهد بود.

در این مورد اگر O مبین تمام دنباله‌های ممکن و $\bar{O}$ مبین بهترین دنبالة سنتز از دیدگاه نوایی (به منظور تولید گفتار طبیعی) باشد، آنگاه موضوع بهینه‌یابی مذکور را می‌توان در قالب رابطة (۳) بیان نمود [۴۱-۴۳]:

$$P(\bar{O} | S_1, \cdots, S_N, CTX_1, \cdots, CTX_N, O) = \arg\max_{O} \left(\prod_{i=1}^{N} P(C(S_i, j) | S_i, CTX_i) \prod_{i=2}^{N} P(C(S_{i-1}, j), C(S_i, \ell) | S_{i-1}, S_i, CTX_i) \right) \qquad (۳)$$

۲-۲- تحقق مدل آمیختار نوایی

در تحقق مدل آمیختاری که در این مقاله برای تولید و انتخاب بهترین نوا ارائه شده است، ابتدا به کمک یک شبکة عصبی بازگشتی تخمین‌هایی از عوامل نوا در سطح هجا[21] به‌دست آمده است. سپس درخت‌های تصمیم‌گیری نوایی را رشد داده [۴۴] تا اطلاعات در مجموعه‌ای از خوشه‌های با بافت یکسان دسته‌بندی شوند. در نهایت نیز مدل مخلوط گوسی برای هر خوشه آموزش داده شده است.

۱-۲-۲- پارامتری ساختن عوامل نوا

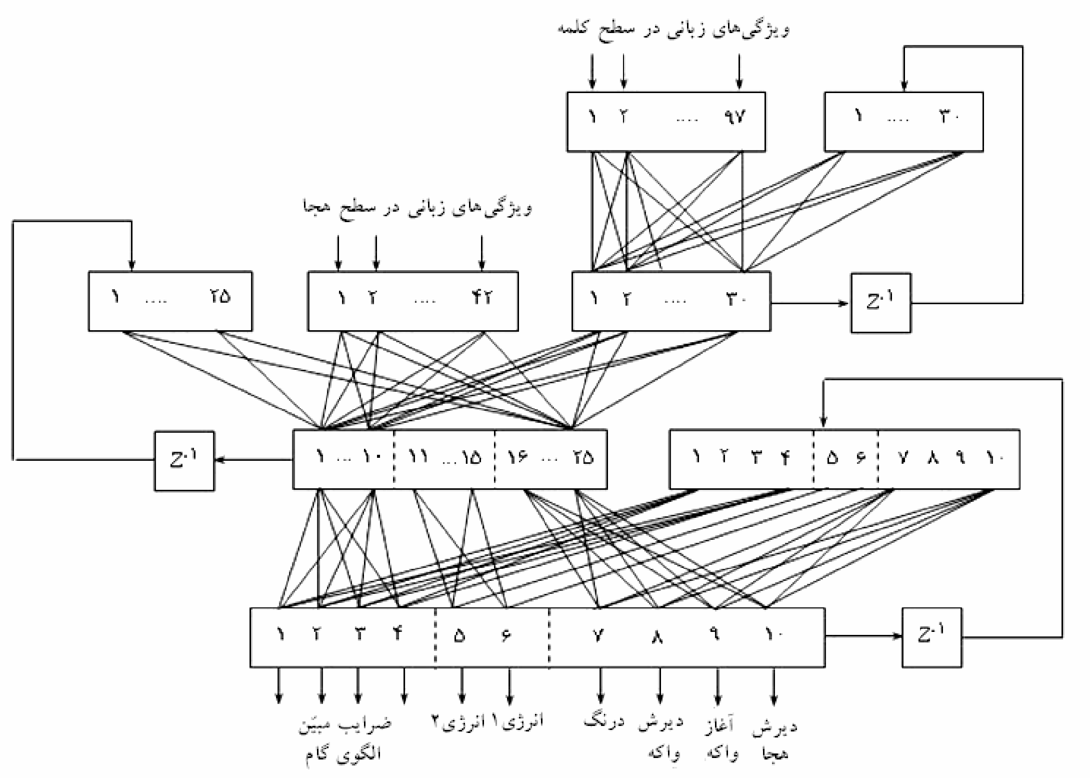
در این قسمت از یک RNN چهار لایه با ساختار ارائه شده در شکل ۱ برای تخمین نوا استفاده شده است [۴۵]. ویژگی‌های ورودی در سطح کلمه به این شبکه عبارت‌اند از: مشخصة دستوری کلمه جاری و بعدی (با توجه به ۴۱ نوع POS پیشنهادی برای زبان فارسی [۶])، طول کلمة جاری و بعدی (برحسب تعداد هجا و بیشینه ۵ هجا در هر کلمه) و نوع علامت ("نقطه"، "ویرگول"، "سؤال"، "تعجب" و "نقل قول") پس از کلمة جاری. ویژگی‌های ورودی در سطح هجا نیز عبارت‌اند از: نوع همخوان[22] اولیة هجای جاری و بعدی (با توجه به دسته‌بندی همخوان‌های زبان فارسی به ۶ دسته با توجه به نحوة ادا [۶])، نوع واکة[23] هجای جاری و بعدی، 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").

---

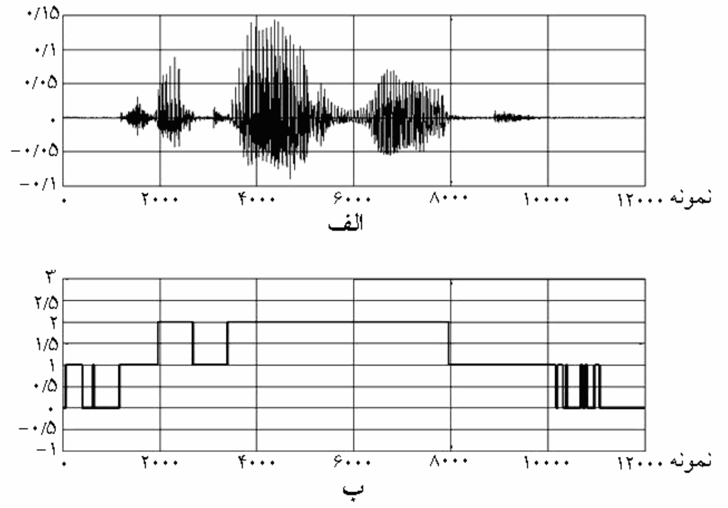
[21] Syllable

[22] **C**onsonant

[23] **V**owel



شکل ۱- ساختار RNN تخمین‌گر پارامترهای نوای گفتار فارسی



شکل ۲- تعیین V/U/S قطعة گفتاري: الف. سيگنال زماني؛ ب. نوع قطعه

چنان‌که در شکل ۱ مشاهده می‌شود، اطلاعات نوایی هجای جاری در لایة خروجی چنین تخمین زده می‌شوند:

الف. چهار پارامتر (ضرایب مراتب صفر تا سوم بسط چندجمله‌ای لژاندر گسسته[24]) مبین الگوی گام [۴۶]؛

ب. دو پارامتر مبین انرژی کوتاه مدت نیمه‌های اول و دوم هجا؛

پ. یک پارامتر مبین میزان "درنگ قبل از هجا" (بر حسب میلی‌ثانیه)؛

ت. سه پارامتر مبین میزان دیرش قطعات "هجا"، "واکة هجا" و "نقطة آغاز قطعة واکدار" (بر حسب میلی‌ثانیه).

تعداد نرون‌ها در لایه‌های مخفی اول و دوم نیز به صورت تجربی به‌ترتیب ۳۰ و ۲۵ در نظر گرفته شده‌اند. برای کاهش پیچیدگی سیستم نیز نرون‌ها در لایة مخفی دوم و خروجی به سه گروه تفکیک و اتصالات به صورت غیرکامل برقرار شده‌اند.

---

[24] Discrete Legendre polynomial expansion

شایان ذکر است برای فراهم نمودن داده‌های آموزشی خودکار برای RNN، تقطیع و نام‌گذاری قطعات گفتار در سطح واج لازم بوده که در ادامه به چگونگی انجام این کار پرداخته می‌شود.

در این مورد روش تقطیع سیگنال گفتار در سطح رویداد[25] [۴۷] و نیز چگونگي تعیین واکداري، بي‌واکي و سکوت این رویدادها بیان مي‌شود. سپس چگونگي برچسب‌زني[26] سیگنال گفتار در سطوح مختلف جمله (کلمه، هجا و واج) بیان خواهد شد. بدین‌ترتیب امکان استخراج خودکار اطلاعات نوا فراهم آمده و داده‌هاي آموزشي براي سیستم مولد نواي گفتار تأمین مي‌شود.

۲-۲-۱-۱- تعیین واکداري/ بي‌واکي/ سکوت

در این عملیات، فرکانس نمونه‌برداري ۱۰kHz، نمونه‌ها ۱۶bit و قطعات ۲۵ msec در نظر گرفته شده‌اند. در ابتدا، براي تعیین واکداري/ بي‌واکي قطعات، اولین ضریب انعکاس[27]، سیگنال مانده[28]، انرژي و ضرایب AR[29] محاسبه مي‌شوند. مقدار انرژي نیمه‌هاي اول و دوم فریم جاري را نیز به‌دست آورده (آنها را "انرژی۱" و "انرژی۲" نامیده) و مقدار "شاخص انرژي" به کمک رابطة (۴) تعیین مي‌گردد [۴۷]:

$$\text{شاخص انرژی} = \sqrt{\text{انرژی ۲} \times \text{انرژی ۱}} \qquad (۴)$$

در صورتي که مقدار اولین ضریب انعکاس از آستانه ۰/۴ و شاخص انرژي مذکور (با فرض ضرب سیگنال در ضریب ((۱۴۰۰۰/max(abs(signal))) از آستانة ۳*۱۰۷ به‌صورت همزمان فراتر روند، قطعه به‌صورت تجربي واکدار در نظر گرفته می‌شود (براي این قطعات، فرکانس گام، دامنه و فرکانس فرمنت‌ها محاسبه مي‌گردند). در غیر این صورت قطعات بي‌واك/ سکوت‌اند که براي تشخیص نوع آنها نیز ابتدا تابع حجم[30] در بازة فرکانسي ۳۰۰Hz تا ۵kHz تعیین مي‌شود:

$$V(i) = \frac{1}{N_i}\sqrt{\sum_{m=17}^{256}\left|H_i(e^{j\pi m/256})\right|} \qquad (۵)$$

که i اندیس فریم جاري، N_i تعداد نمونه‌هاي فریم و $H_i(z)$ تابع انتقال با رابطة (۶) است:

$$H_i(z) = \frac{G(i)}{1 + a_1 z^{-1} + \ldots + a_{13} z^{-13}} \qquad (۶)$$

G(i) نیز بهره بوده و از رابطة (۷) به‌دست مي‌آید:

$$G(i) = \sqrt{\sum_{n=<250>} r^2(n)} \qquad (۷)$$

r(n)، سیگنال ماندة فریم جاري است. با داشتن مقدار بهره و ضرایب AR، توان چند میلي‌ثانیه ابتدایي سیگنال (که سکوت است)، به‌دست آمده و مقدار میانگین (m) و انحراف معیار (σ) آن تعیین مي‌گردد. در این راستا، آستانه‌اي به صورت رابطة (۸) تعیین و در نظر گرفته مي‌شود:

$$T_{U/S} = m + K.\sigma \qquad (۸)$$

مقدار K نیز به صورت تجربي و در جریان اداي ۱۰۰ جمله (با نسبت سیگنال به نویز برابر ۳۰ dB) برابر ۲ در نظر گرفته شد. در صورتي که ۲۰log۱۰V(i) فریم مورد نظر بیش از $T_{U/S}$ باشد، آن را بي‌واك و در غیر این صورت آن را به عنوان سکوت در نظر مي‌گیریم. البته بدیهی است که در جریان این پیش‌پردازش ساده، امکان بروز اشتباه (به‌ویژه در تشخیص قطعات بي‌واک و سکوت) وجود دارد. اما با توجه به مراحل پردازش تکمیلی ارائه شده در ادامه، خطای موجود کمینه می‌شود. شکل ۲ نوع قطعات (V/U/S) را براي عبارت "چطور بود؟".

نشان مي‌دهد (V/U/S متناظر با سطوح ۲/۱/۰ است).

فرکانس‌هاي گام و فرمنت نیز با توجه به امکانات ارائه شده در مرجع [۴۷] تعیین شده‌اند.

۲-۲-۱-۲- تقطیع رویدادها

گفتار را در سطح جزئي تحت عنوان رویداد، تقطیع کرده تا پس از تعیین ارتباط بین رویدادها، سیگنال در سطح واج قطعه‌بندي شود. براي تعیین مرز این رویدادها نیز از بررسی تغییرات مشخصات زیر به عنوان سه روش مکمل استفاده شده است:

الف. وابستگي طیفي کوتاه- مدت سیگنال؛

ب. مشخصات آکوستیکي؛

ج. دامنه و فرکانس فرمنت‌ها.

---

[25] Event
[26] Labeling
[27] Reflection coefficient
[28] Residue
[29] **A**uto-**R**egressive
[30] Volume function

البته روش اول در تعيين مرز واج‌هاي انفجاري با مشكل مواجه است. در روش دوم نيز اساس كار بر تقسيم‌بندي گفتار بر حسب انواع رسا[۳۱]، واكه، همخوان واكدار[۳۲]، خيشومي[۳۳]، شبه‌واكه[۳۴] و سايشي واكدار[۳۵] است. در اين روش سيگنال گفتار برحسب شش نوع مذكور، دسته‌بندي و امتيازدهي مي‌شود. بدين‌ترتيب كه تابع حجم رابطة (۵) با حدود سيگمای A و B (مبين فرکانس‌های قطع پایین و بالاي يك فيلتر ميان‌گذر) در نظر گرفته مي‌شود. $H_i(z)$ نيز همان است كه در رابطة (۶) معرفي شد. سپس توابع حجم فرکانس پایین (LFV)[۳۶] و حجم فرکانس بالا (HFV)[۳۷] به ازای مقادير متفاوتي از G، A و B محاسبه و نسبت آنها تشکیل می‌شود:

$$R(i) = \frac{LFV\ (i)}{HFV\ (i)} \qquad (۹)$$

پس از اعمال يك فيلتر ميانة مرتبة ۵ به R(i)، امتياز برخي از دسته‌ها به عنوان نمونه به‌شكل رابطة (۱۰) محاسبه مي‌شود:

$$Score(i) = \begin{cases} 1 & ;\ R(i) \ge T_{up} \\ 0 & ;\ R(i) < T_{low} \\ \dfrac{R(i) - T_{low}}{T_{up} - T_{low}} & ;\ T_{low} \le R(i) < T_{up} \end{cases} \qquad (۱۰)$$

مقادير به‌كار رفته براي محاسبة امتياز مذكور براي اصوات همخوان واکدار، قطعات با فرکانس فرمنت خیلی پایین[۳۸] (یعنی حدود ۱۵۰Hz که برای محاسبة امتیاز شبه‌واکه‌ها استفاده می‌شود) براساس روند مذكور، به عنوان نمونه در جدول ۱ آمده است.

براي دسته‌هاي ديگر نيز امتياز با قدري تغيير در روابط قابل محاسبه است [۴۷،۴۸]. شکل ۳ نتايج اين امتيازبندي را براي قطعات غيرواكه گفتار موضوع شکل ۲ نشان مي‌دهد.

روش سوم نيز در مواقعي كه دو روش قبلي نتوانند مرز بين واج‌های واكدار را (به هنگام اداي گفتار پيوسته سريع) آشكار كنند، به‌كار گرفته مي‌شود.

جدول ۱- مقادير پارامترهاي به‌كاررفته در تعيين امتياز دسته‌هاي آكوستيكي همخوان، واكدار و قطعات با فرکانس فرمنت خیلی کم (به عنوان دستة کمکی)

نوع دسته		همخوان واكدار	قطعه با فرکانس فرمنت خیلی کم
LFV	G	۱	۱
	A	۱	۱
	B	۵۱	۳۳
HFV	G	۱	۱
	A	۵۲	۳۴
	B	۲۵۵	۲۵۵
T_{up}		۱۸	۳۰
T_{low}		۸	۱۰
نام اختصاری امتیاز		VCS	VBS

۲-۲-۱-۳- نامگذاري خودكار قطعات مختلف گفتار

جهت تشخيص واج‌ها و نيز تعيين مرزهاي هجاها و كلمات در جملات اداشده، مراحل كاري زير انجام شده است:

الف. تشخيص واج‌هاي واكدار؛

ب. تشخيص واج‌هاي بي‌واك و سكوت؛

ج. نامگذاري واج‌هاي واكدار با توجه به محل واكه‌ها؛

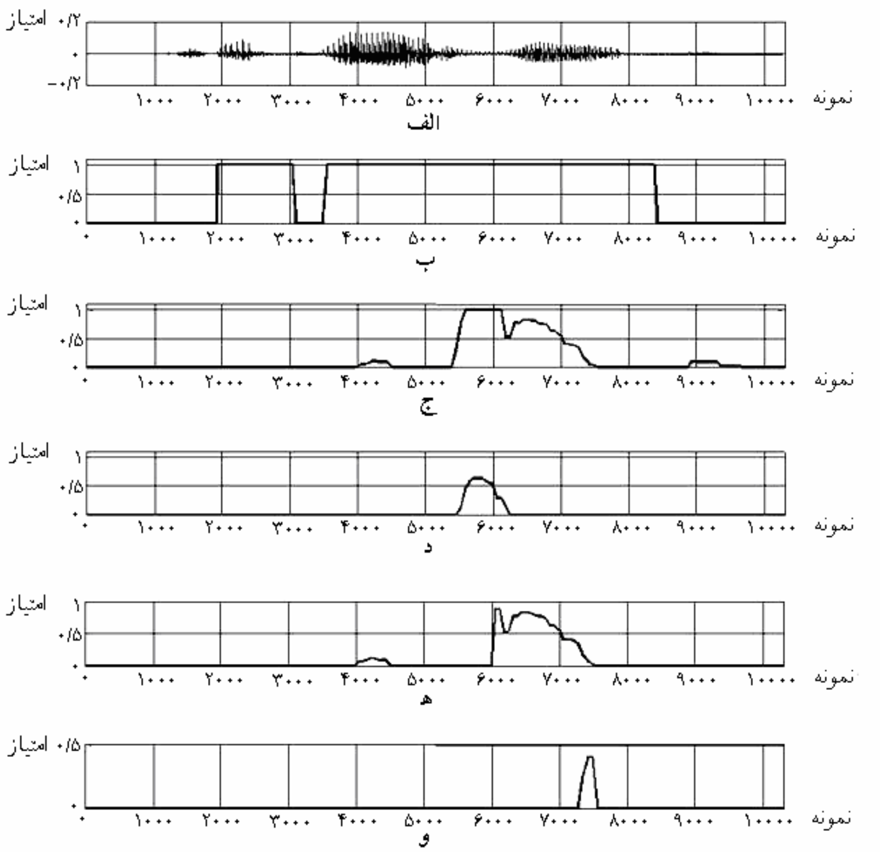
د. نامگذاري تمام واج‌هاي جمله.

براي تشخيص واج‌هاي واكدار، از يك شبكة عصبي پرسپترون چندلايه (MLP)[۳۹] با ۹ گره در لاية ورودي (شامل اطلاعات فركانسي سه فرمنت اول قطعة جاري و دو قطعة مجاور آن)، دو لاية مياني به‌ترتيب با ۲۰ و ۳۰ گره و ۲۹ گره در لاية خروجي (به تعداد واج‌هاي زبان فارسي) استفاده شده است.

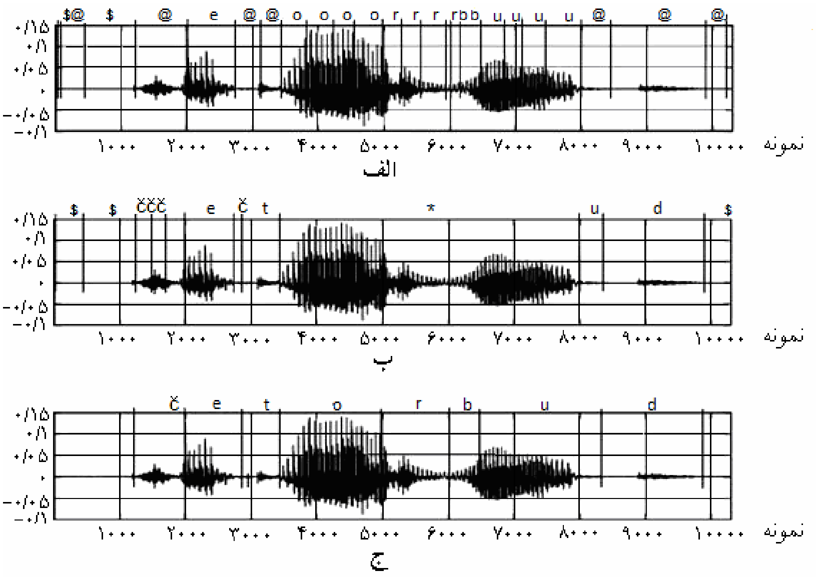
براي تشخيص واج‌هاي بي‌واك/ سكوت نيز از MLP ديگري با ۹ گره در لاية ورودي (شامل انرژي سيگنال در فركانس‌هاي بيش از ۲kHz، انرژي سيگنال در فركانس‌هاي كمتر از ۵kHz و نرخ عبور از صفر مربوط به قطعة جاري و دو قطعة مجاور آن)، دو لاية مياني به‌ترتيب با ۲۰ و ۳۰ گره و نيز ۳۰ گره در لاية خروجي (به تعداد واج‌هاي زبان فارسي به علاوة سكوت) استفاده شده است.

---

[31] Sonorant
[32] Voiced consonant
[33] Nasal
[34] Semivowel
[35] Voiced fricative
[36] **L**ow **F**requency **V**olume function
[37] **H**igh **F**requency **V**olume function
[38] Voice bar
[39] **M**ulti-**L**ayer **P**erceptron



شکل ۳- امتيازبندي آکوستيکي قطعات گفتاري: الف) قطعة گفتاری؛ ب) امتياز رسا؛ ج) امتياز همخوان واکدار؛ د) امتياز خيشومی؛ ه.) امتياز شبه‌واکه؛ و) امتياز سايشی واکدار



شکل ۴- نامگذاري قطعات گفتاري عبارت"چطوربود؟": الف) قطعات واکدار [$ و @: قطعات نامگذاري نشده]؛ ب) قطعات بي‌واك [$: سکوت و *: قطعات واکدار نامگذاري شده]؛ ج) نامگذاري واجي قطعات

با استفاده از خروجی شبکهٔ عصبي مذکور و نیز بـا توجـه بـه متن ورودي، نامگذاري واج‌هاي واكدار صورت مي‌گيرد. براي جلوگيري از انتشار خطا نيز ابتدا محل يك قطعه از هـر واكـه تشخيص داده مي‌شود (بدين‌ترتيب خطاي نامگـذاري بـين دو واكه محدود مي‌شود) و براي تعيين محـل يـك قطعـه از هـر واكه نيز در منحني بهره (رابطهٔ ۷) بيشينه‌هاي محلـي جـستجو می‌شوند. تعداد اين بيشينه‌ها، به‌طـور معمـول بـيش از تعـداد واكه‌هاست. براي حذف بيشينه‌هاي اضافي نيز ۲۵ معيار به‌كـار گرفته شده‌اند كه به نمونه‌هايي از آنها در زير اشـاره مـی‌شـود (كه به صورت تكي يا تركيبي اعمال مي‌شوند):

الف. نسبت نرخ عبور از صفر بهنجارشده به طول فريم، كمتـر از ۰/۳۳ باشد؛

ب. مقدار انرژي مؤلفه‌هاي فركانسي كمتر از ۲/۵kHz، بيـشتر از مؤلفه‌هاي فركانسي ۲/۵kHz تا ۵ kHz باشد؛

ج. براي دو قطعهٔ مجاور كه واكـه تـشخيص داده شـده‌انـد و فاصلهٔ آنها كمتر از۱۰۰msec است، همبستگي طيفـي كمتـر از ۰/۸۵ باشد.

دقت اين الگوريتم ۹۷/۲% است و براي رسيدن بـه دقـت بيشتر، روش دستي اعمال شـده اسـت. در ايـن راسـتا، نتـايج نامگذاري قطعات واكدار، بي‌واك و واجي قطعات در شكل ۴ نشان داده شده است.

براي تعيين مرز هجاها نيز با توجه به‌حالت‌هـاي مختلـف هجايي در زبان فارسي (CV، CVC و CVCC)، با پيدا كـردن محل واكه‌ها و در نظر گرفتن واج قبل از آن مي‌تـوان براحتـي مرزهاي هجا را مشخص نمود. با توجه به تقطيع و نامگـذاري خودكار قطعات مي‌تـوان پارامترهـاي نـوايي مربـوط (الگـوي فركانس گام، ديرش، شدت و درنگ) را تعيين و بدين‌ترتيـب داده‌هاي آموزشي را به‌صـورت خودكـار بـراي سيـستم مولـد نواي گفتار فراهم آورد.

۲-۲-۲- تشکیل درخت‌های نوایی

دو نوع درخت نوايی که یکی متناظر با مـدل نـوايی هـدف و ديگری متناظر با مدل نوايی گذار اسـت، رشـد داده شـده‌انـد. درخت‌های هدف نوايی، خوشه‌هايی با بافت يکسان از مقادير هدف را توليد می‌کنند و درخت‌های گذار نـوايی نيـز مقـادير گذار مربوط را توليد می‌نمايند.

در همين راستا، دو نوع درخت هدف در اين تحقيق رشد داده شده‌اند:

الف. درخت هدف "گام + ديـرش" کـه در آن ($p۱_n$، $p۲_n$، $p۳_n$، $p۴_n$، $d۱_n$، $d۲_n$، $d۳_n$) بردار ويژگی است و مقـادير p۱ تـا p۴ همان خروجی‌های اول تـا چهـارم در RNN (کـه مربـوط بـه منحنی گام‌اند) و مقادير d۱ تا d۳ همـان خروجـی‌هـای ۷، ۸، ۱۰ درRNN (به‌ترتيـب مبـين ديـرش‌هـای "هجـا"، "واکـه" و "درنگ قبل از هجای n اُم") می‌باشند؛

ب. درخت هدف "انرژی" که در آن بردار ويژگی (e۱n, e۲n) بوده، و مبين انرژی کوتاه مدت نيمه‌های اول و دوم هجـای n اُم‌اند.

به همين ترتيب سه نوع درخت گذار نيـز مـورد اسـتفاده قرار گرفته‌اند:

الـــف. درخـــت گـــذار "گـــام" کـــه در آن بـــردار ويژگـــی ($p۱_n$-$p۱_{n-۱}$,…, $p۴_n$-$p۴_{n-۱}$) و n وn-۱ نـــشانگر دو هجـــای متوالی‌اند؛

ب. درخت گذار "ديرش" که ($d۱_n$-$d۱_{n-۱}$,…, $d۳_n$-$d۳_{n-۱}$) در آن بردار ويژگی است؛

ج. درخـــت گـــذار "انـــرژی" کـــه در آن بـــردار ويژگـــی ($e۱_n$-$e۱_{n-۱}$, $e۲_n$ - $e۲_{n-۱}$) است.

لازم به ذکر است که درخت‌های تصميم‌گيـری کـه بـرای دسته‌بندی اطلاعات در مجموعـه‌ای از خوشـه‌هـای بـا بافـت يکسان، در انواع "درخـت هـدف" و "درخـت گـذار" سـاخته شده‌اند، دودويی[40] بوده و لذا خروجی آنها صفر يا يک است.

۳-۲-۲- آموزش مدل مخلوط گوسی

پس از اينکه درخت‌ها ساخته و خوشه‌ها توليد شـدند، مـدل مخلوط گوسی برای هر خوشه آمـوزش داده مـی‌شـود. مـدل مخلوط اين امکان را فراهم می‌آورد که توزيع احتمالاتی متغير Z به صورت مجموع مخلوط Q مؤلفه مدل‌سازی شـود. تـابع توزيع مذکور نيز به صورت رابطهٔ (۱۱) است:

$$P_{GMM}(Z;\alpha,\mu,\Sigma)=\sum_{q=1}^{Q}\alpha_q N(Z;\mu_q,\Sigma_q) \qquad (۱۱)$$

که در آن $\sum_{q=1}^{Q}\alpha_q=1(\alpha_q\geq 0)$ و α_q مبين احتمالات اوليـهٔ Z است که به‌وسيلهٔ مؤلفهٔ q توليد شـود. $N(Z;\mu,\Sigma)$ نيـز مبـين

---

[40] Binary

توزیع نرمال n - بعدی با بردار میانگین μ و ماتریس کوواریانس Σ است، که در رابطة (۱۲) آمده است:

$$N(Z;\mu,\Sigma)=\frac{1}{(2\pi)^{n/2}\sqrt{|\Sigma|}}\exp\left[-\frac{1}{2}(Z-\mu)^{T}\Sigma^{-1}(Z-\mu)\right] \qquad (۱۲)$$

پارامترهای GMM ($\{\alpha,\mu,\Sigma\}$) با به‌کارگیري الگوریتم EM[۴۱] تخمین زده شده‌اند [۴۹]. در این مورد لازم به توضیح است که ساختار GMM های به‌کار گرفته‌شده از نوع قطری[۴۲] و تعداد آنها نیز به اندازة برگ‌های درخت تصمیم‌گیری است. در این پژوهش با توجه به داده‌های آزمون و در جریان سه آزمایش مختلف، تعداد برگ‌ها ۷۴۳، ۴۰۷ و ۲۱۰ به‌دست آمد. بدین‌ترتیب به همین تعداد GMM آموزش داده شده است. ساختار بردار آموزش این GMM ها نیز با توجه به دو نوع درخت هدف و نیز سه نوع درخت گذار (بیان‌شده در بند ۲-۲-۲) تعیین می‌شود. به عنوان نمونه از بردار ($p۱_n-p۱_{n-۱}, p۲_n-p۲_{n-۱}, p۳_n-p۳_{n-۱}, p۴_n-p۴_{n-۱}$) برای آموزش GMM های متناظر با برگ‌های درخت گذار گام استفاده می‌شود.

تعداد مؤلفه‌ها در مدل مخلوط گوسی (Q) نیز، همان‌گونه که در نتایج ارائه شده در ادامه مشهود است، از ۱ تا ۵ مؤلفه در آزمایش‌های مختلف (و بیشینه ۵ برای تمامی GMM های متناظر با برگ‌های درخت‌های نوایی مختلف) انتخاب شده است.

در زمان اجرا، اطلاعات بافت به‌وسیلة پردازشگر زبان طبیعی تولید می‌شوند [۶] و سپس هر درخت با پرسش از متن برای یافتن GMM وابسته به بافت پیمایش می‌شود. این GMM ها برای محاسبة هزینة هدف و گذار در فرایند جستجو به‌کار گرفته می‌شوند.

در ساختارهای هجایی با آغاز بي‌واك- واكه (UV) ، اولین نقطه از منحنی گام در به‌کارگیری مدل‌های آماری و درخت‌های تصمیم‌گیری نامشخص بوده و از رگرسیون مخلوط گوسی برای پیش‌بینی مقادیر مجهول استفاده می‌کنند [۴۱،۳۱] اما مشکل یادشده در این پژوهش با به‌کارگیری RNN برای تخمین کل منحنی گام هجا (به کمک چهار ضریب بسط چندجمله‌ای لژاندر گسسته) برطرف شده است.

۲-۳- سنتز مبتنی بر انتخاب قطعات گفتاری

سه رویکرد اصلی برای سنتز مبتنی بر اتصال قطعات گفتاری[۴۳] وجود دارد: سنتز در حوزه‌های کاربردی مشخص [۵۰]، سنتز دوواجی‌ها [۵۱-۵۳] و سنتز مبتنی بر انتخاب واحد [۵۴-۵۷]. در سنتز مبتنی بر انتخاب واحد، از دادگان سنتز حجیم‌تری (متشکل از واحدهای هجایی، کلمات، عبارات یا جملات) استفاده شده و گفتار تولیدی نیز طبیعی‌تر است. در این پژوهش بر خلاف رویکرد سنتز اتخاذشده در مرجع [۶] (سنتز دوواجی‌ها به روش HNM)، از واحد هجا در سنتز مبتنی بر انتخاب واحد بهره گرفته شد. دادگان هجایی سنتز از روی ۵۰۰۰ جملة اداشده در استودیوی سازمان صدا و سیمای جمهوری اسلامی ایران با تنوع کافی واجی و دستوری [۵۸] و تقطیع و نامگذاری واجی خودکار در سطح هجا تشکیل شده است [۴۸].

در راستای ذخیره‌سازی پارامتری و سنتز قطعات هجایی نیز از مدل HNM1 استفاده شده است. در این راستا، الگوریتم‌های تعیین دامنه و فاز هارمونیک‌ها [۵۹]، فرکانس واکداری بیشینه (MVF)[۴۴] [۶۰]، تصحیح فاز هر فریم آنالیز [۶۱]، تعیین پارامترهای HNM حول نقاط اتصال قطعات گفتار، اصلاح نوا و تعیین لحظات زمانی سنتز مربوط [۶۲] و تخمین مقادیر دامنه و فاز هارمونیک‌ها با فرکانس گام جدید [۶۳]، هموارسازی[۴۵] ناپیوستگی‌های حاصل از عدم انطباق عوامل فاز، گام و پوش طیفی در محل اتصال هجاها و سنتز قطعات گفتاری [۶۴] پیاده‌سازی و به‌کار گرفته شده‌اند.

۳- نتایج و بحث

در این بخش، عملکرد RNN در پارامتری‌ساختن عوامل نوا، رفتار مدل نوایی احتمالاتی با توجه به تعداد برگ‌ها در درخت تصمیم‌گیری و مؤلفه‌ها در هر GMM و در نهایت عملکرد کل سیستم از نظر کیفیت گفتار تولیدی مورد ارزیابی قرار گرفته است.

از ۵۰۰۰ جمله‌ای که به‌وسیلة یک گویندة مرد با سرعت میانگین ۴ هجا در ثانیه ادا شد، برای آموزش RNN استفاده

---

[41] Expectation – Maximization

[42] Diagonal

[43] Concatenative synthesis

[44] Maximum Voiced Frequency

[45] Smoothing

کرده و با بهره‌گیری از "جعبه‌ابزار[۴۶] شبکه‌های عصبی"در نرم‌افزار MATLAB[۴۷]، ساختار ارائه شده در شکل ۱ شبیه‌سازی شده است. برای ارزیابی عملکرد سیستم نیز ۳۰۰ جمله خارج از مجموعهٔ آموزشی انتخاب و پس از اتمام فاز آموزش، RMSE[۴۸] مربوط به مقادیر پارامترهای ارائه شده به‌وسیلهٔ RNN در مورد عوامل نوا به‌دست آمد (جدول۲).

نمایی از الگوی گام واقعی مربوط به عبارت "کجا آمدی؟" و الگوی گام حاصل از بازسازی به کمک ضرایب مراتب صفر تا سوم بسط چندجمله‌ای لژاندر گسسته (یا به عبارتی تخمین مربوط به‌وسیلهٔ RNN) نیز در شکل ۵ نشان داده شده است. در این شکل منحنی پررنگ‌تر مربوط به الگوی گام تخمینی به‌وسیلهٔ RNN است.

عامال‌های کنترل‌کنندهٔ مختلفی نیز در مدل نوایی آماری وجود دارند. در این بین تعداد برگ‌ها در درخت‌های تصمیم‌گیری و تعداد مؤلفه‌ها در هر GMM در این ارزیابی مد نظر قرار داده شد. بر این اساس، مدل‌های گام مختلفی با ترکیب‌های متفاوت از این دو عدد ساخته شده و روی دو مجموعه از داده‌ها (یکی برای آموزش و دیگری برای آزمون) مورد آزمون قرار گرفته‌اند. با توجه به این‌که هزینه[۴۹] یک رکورد در دادگان به صورت "منهای لگاریتم تابع شباهت[۵۰] ارائه شده به‌وسیلهٔ مدل احتمالاتی" به‌دست می‌آید، میانگین هزینه برای هر مدل از دو مجموعه داده محاسبه شده است.

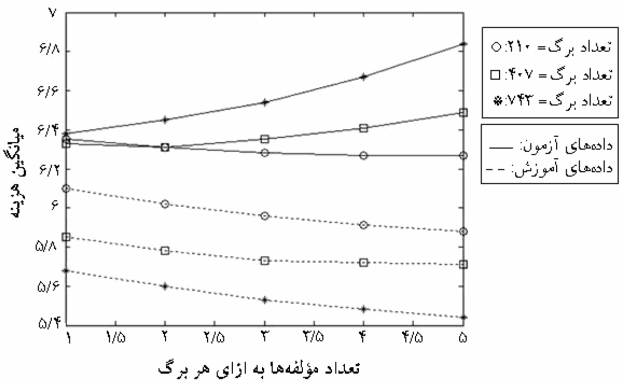
برای مجموعهٔ داده‌های آموزشی، هر چه تعداد برگ‌ها و مؤلفه‌ها بیشتر باشد، میانگین هزینه کمتر خواهد بود. اما برای مجموعهٔ داده‌های آزمون، مقدار کمینه برای میانگین هزینه با ترکیب مناسبی از این دو پارامتر به‌دست مي‌آید.

در شکل ۶ میانگین هزینه برای مدل‌های گام مختلف مربوط به مجموعهٔ داده‌های آموزشی و آزمون و به ازای تعداد برگ‌های گوناگون ترسیم شده است. همان‌گونه که مشاهده می‌شود، وقتی تعداد برگ‌ها و مؤلفه‌ها زیاد باشد، مدل دچار آموزش بیش از حد[۵۱] می‌شود.

جدول ۲- RMSE پنج پارامتر نوای تولیدشده به‌وسیلهٔ RNN

فرکانس گام(Hz)	انرژی (dB)	دیرش هجا (msec)	دیرش واکه (msec)	دیرش درنگ (msec)
۱۸/۷	۲/۱	۴۰/۸	۳۶/۰	۳۱/۲

شکل ۵- مقایسهٔ الگوی گام واقعی و تخمین زده‌شده به‌وسیلهٔ RNN مربوط به عبارت "کجا آمدی؟"



شکل۶- میانگین هزینه مدل‌های مختلف گام مربوط به مجموعهٔ داده‌های آموزشی و آزمون

برای ارزیابی عملکرد کل سیستم که مبین کارایی مدل آمیختار نوایی ارائه شده نیز می‌باشد، از استاندارد ITU-T P.85 استفاده شده است [۶۵]. براساس این استاندارد، ملاک‌های ارزیابی به‌کاررفته (با قیاس MOS[۵۲]) عبارتند از:

الف. نظر کلی[۵۳] (دربارهٔ کیفیت صدا)؛
ب. تلاش شنیداری[۵۴] (برای درک پیام)؛
ج. میزان درک[۵۵] (مشکل در فهم کلمات)؛
د. نحوهٔ ادا[۵۶] (میزان تمایز بین صداها)؛
ه. کیفیت تلفظ[۵۷]؛
و. خوشایندی[۵۸] صدا.

---

[46] Toolbox
[47] MATrix Laboratory
[48] Root Mean Squared Error
[49] Cost
[50] Likelihood
[51] Overtraining
[52] **M**ean **O**pinion **S**core
[53] Overall impression
[54] Listening effort
[55] Comprehension
[56] Articulation
[57] Pronunciation
[58] Pleasantness

جدول ۳- نتایج ارزیابي عملکرد سیستم TTS ارائه شده برای زبان فارسی(FTTSv.2 وFTTSv.1) و چند سیستم نمونة جهانی

ملاك ارزيابي	نام سیستم							
	EL	LT	AK	RS	SS	ATT	FTTSv.1	FTTSv.2
نظر کلی	۱/۷	۲/۲	۲/۳	۳/۱	۳/۶	۳/۷	۳/۰	۳/۵
تلاش شنیداری	۲/۳	۲/۸	۲/۸	۳/۴	۳/۷	۳/۷	۳/۲	۳/۶
میزان درك	۲/۵	۳/۰	۳/۰	۳/۶	۳/۹	۳/۹	۳/۱	۳/۸
نحوة ادا	۲/۲	۲/۷	۲/۷	۳/۳	۳/۷	۳/۸	۳/۲	۳/۸
کیفیت تلفظ	۲/۰	۲/۶	۲/۶	۳/۲	۳/۵	۳/۳	۲/۷	۳/۴
خوشايندي صدا	۱/۲	۲/۱	۲/۵	۳/۱	۳/۶	۳/۷	۲/۸	۳/۶
برآورد کلي (میانگین ملاك‌هاي فوق)	۲/۰	۲/۶	۲/۷	۳/۳	۳/۷	۳/۷	۳/۰	۳/۶

از بیست شنونده (با سنین ۱۵ تا ۴۵ سال) برای اظهارنظر در مورد ۵۰ جمله (با حداقل دیرش ۱۰ ثانیه) استفاده شده و به دلیل تعدد ملاک‌ها، هر جمله سنتزشده دوبار پخش گردید تا در بار اول در مورد سه ملاک اول و در بار دوم در مورد سه ملاک دوم اظهار نظر شود.

نتایج به‌دست‌آمده از سیستم حاضر و نیز سیستم توسعه‌یافته در پژوهش قبلی (که به‌ترتیب FTTSv.2 وFTTSv.1 نامیده شد) و نیز ۶ سیستم TTS خارجی با نام‌های اختصاری انتخابی EL[59]، LT[60]، AK[61]، RS[62]، SS[63] و ATT[64] (که نتایج ارزیابی آنها در شرایط ارائه شده در استانداردP.85 ازITU-T در مرجع [۶۶] گزارش شده بود)، در جدول ۳ آمده است.

شایان ذکر است که FTTSv.1 ، که نتایج عملکرد آن در مرجع [۶] نیز گزارش شد، دارای ۱۲ واحد در خروجی و ۳۰ واحد در لایة مخفی دوم RNN بود که این تعداد در این تحقیق به‌ترتیب با ۱۰ و ۲۵ جایگزین شده است.

۴- نتیجه‌گیری

در این مقاله، با توجه به در اختیار داشتن سیستم پردازشگر زبان طبیعی [۶]، اصلاحاتی در بخش مولد نوای گفتار و نیز سنتز (درمقایسه با ویرایش اول این سیستم) انجام گردید و نتایج آن ارائه شد. بدین ترتیب که از یک شبکة عصبی بازگشتی برای ارائة پارامترهای الگوی فرکانس گام (در قالب ضرایب مراتب صفر تا سوم بسط چندجمله‌ای لژاندر گسسته)، دیرش (هجا، واکة درون هجا و درنگ قبل از هجا)، نقطة آغاز قطعة واکدار درون هجا و انرژی کوتاه- مدت نیمه‌های اول و دوم هجا استفاده شد.

با به‌کارگیری پارامترهای مذکور، دو نوع درخت نوایی (مدل هدف و مدل گذار) رشد داده شد تا اطلاعات در مجموعه‌ای از خوشه‌های با بافت یکسان دسته‌بندی شوند. برای هر خوشه نیز مدل مخلوط گوسی آموزش داده شد.

در زمان اجرا، اطلاعات بافت به‌وسیلة سیستم NLP تولید شده و سپس هر درخت با پرسش از متن برای یافتن GMM وابسته به بافت پیمایش شد. این GMMها برای محاسبة هزینة هدف و گذار در فرایند جستجو به‌کار گرفته شدند.

در نهایت نیز با ارائة چارچوبی برای انتخاب واحد بهینة گفتاری، امکان انتخاب بهترین دنبالة سنتز از دیدگاه نوایی در یک سنتزکنندة مبتنی بر اتصال قطعات گفتاری با پارامترهای HNM فراهم آمد.

به‌منظور ارزیابی عملکرد بخش‌های مختلف سیستم نیز RMSE پارامترهای نوای ارائه شده به‌وسیلة RNN و میانگین هزینه به ازای تعداد مختلف برگ‌ها در درخت تصمیم‌گیری و نیز تعداد مؤلفه‌ها در GMM، برای مجموعة داده‌های آموزشی و آزمون به‌دست آمد. ارزیابی عملکرد کلی سیستم نیز براساس شش ملاک پیشنهادی در استاندارد P.85 از ITU-T انجام شده و MOS معادل ۳/۶ حاصل گردید.

---

[59] **El**an Informatique
[60] **L**ucent **T**echnology
[61] **A**culab
[62] **R**eal**S**peak
[63] **S**peechworks **S**peechify
[64] **A**merican **T**elephone and **T**elegraph

سپاسگزاری

توسعة نسخة اول سیستم تبدیل متن به گفتار در این مقاله با حمایت حوزة معاونت پژوهشی دانشگاه آزاد اسلامی و نیز همفکری آقایان دکتر رضا صفوی نائینی، دکتر علی فرخی، دکتر داوود غرویان، دکتر کامبیز بدیع، مهندس مجید نصیرزاده و مهندس علی دفتریان انجام شده است. اکنون که نسخة دوم این سیستم آماده و در مقالة حاضر گزارش شده است، از تمام ایشان قدردانی می‌شود.

مراجع

[۱] بی‌جن‌خان محمود؛ بازنمایی واجی و آوایی زبان فارسی و کاربرد آن در بازشناسی خودکار گفتار؛ پایان‌نامة دکتری، دانشکدة ادبیات و علوم انسانی، دانشگاه تهران، ۱۳۷۴.

[۲] سیدصالحی سیدعلی؛ بازشناخت گفتار پیوسته فارسی با استفاده از مدل عملکردی مغز انسان در درک گفتار؛ پایان‌نامه دکتری، دانشکدة فنی و مهندسی، دانشگاه تربیت مدرس، ۱۳۷۴.

[3] Sheikhan M., Tebyani M., Lotfizad M., Continuous speech recognition and syntactic processing in Iranian Farsi language; International Journal of Speech Technology 1997; 1:135-141.

[۴] حافظی محمدمهدی، ثامتی حسین، منصوری نیلوفر، منتظری نیلوفر، بحرانی محمد، موثق حامد؛ ارائة یک مدل دستوری برای بهبود دقت سیستم‌های بازشناسی گفتار پیوستة فارسی؛ مجموعة مقالات دومین کارگاه پژوهشی زبان فارسی و رایانه، دانشگاه تهران، ۱۳۸۵؛ ۸۰-۹۱.

[۵] اسلامی محرم؛ شناخت نوای گفتار زبان فارسی و کاربرد آن در بازسازی و بازشناسی رایانه‌ای گفتار؛ پایان‌نامة دکتری، دانشکدة ادبیات و علوم انسانی، دانشگاه تهران، ۱۳۷۹.

[۶] شیخان منصور، نصیرزاده مجید، دفتریان علی؛ طراحی و پیاده‌سازی سیستم تبدیل متن به گفتار طبیعی برای زبان فارسی؛ نشریة علمی- پژوهشی دانشکدة مهندسی دانشگاه فردوسی مشهد، ۱۳۸۴؛ ۱۷(۲): ۳۱-۴۸.

[۷] شریفی آتشگاه مسعود؛ بررسی نوایی و آوایی واژه‌بست‌های فارسی؛ مجموعة مقالات دومین کارگاه پژوهشی زبان فارسی و رایانه، دانشگاه تهران، ۱۳۸۵؛ ۳۱۰-۳۲۱.

[۸] همایون‌پور محمدمهدی، نم‌نبات مجید؛ فارس بیان: گامی اساسی در سنتز گفتار فارسی؛ مجموعة مقالات دومین کارگاه پژوهشی زبان فارسی و رایانه، دانشگاه تهران، ۱۳۸۵؛ ۳۶۵-۳۷۵.

[۹] کوچاری عباس، نم‌نبات مجید، رحیمی سعید، قاسمی‌زاده بهرنگ، سالاریان مرتضی؛ مدل کردن کشش زمانی واج برای سیستم تبدیل متن به گفتار فارسی به کمک شبکة عصبی؛ مجموعة مقالات دومین کارگاه پژوهشی زبان فارسی و رایانه، دانشگاه تهران، ۱۳۸۵؛ ۳۵۹-۳۶۶.

[10] Sheikhan M., Tebyani M., Lotfizad M., Using symbolic and connectionist approaches to automate editing Persian sentences syntacticly; Proceedings of International Conference on Intelligent and Cognitive Systems 1996; 250-253.

[11] Yamashita Y., Ishida T., Stochastic F0 contour model based on the clustering of F0 shapes of a syntactic unit; Proceedings of European Conference on Speech Communication and Technology 2001; 533-537.

[12] Buhmann J., Marten J.P., Macken L., Van Coile B., Intonation modeling for the synthesis of structured documents; Proceedings of International Conference on Spoken Language Processing 2002; 2089-2092.

[13] Payo V.C., Mancebo D.E., A strategy to solve data scarcity problems in corpus based intonation modeling; Proceedings of International Conference on Acoustics, Speech, and Signal Processing 2004; 665-668.

[14] Aguero P.D., Bonafonte A., Consistent estimation of Fujisaki's intonation model parameters; Proceedings of International Conference on Speech and Computer 2005; 297-300.

[15] Ishi C.T., Ishiguro H., Hagita N., Automatic extraction of paralinguistic information using prosodic features related to F0, duration and voice quality; Speech Communication 2008; 50:531-543.

[16] Smith C.L., Modeling durational variability in eading aloud a connected text; Proceedings of International Conference on Spoken Language Processing 2002; 1769-1772.

[17] Eichner M., Wolff M., Hoffmann R., Improved duration control for speech synthesis using a multigram language model; Proceedings of International Conference on Acoustics, Speech, and Signal Processing 2002; 417-420.

[18] Werner S, Wolff M, Eichner M, Hoffmann R; Modeling pronunciation variation for spontaneous speech synthesis; Proceedings of International Conference on Acoustics, Speech, and Signal Processing 2004; 673-676.

[19] Ariu M., Masuko T., Tanaka S., Kawamura A., Speech recognition using syllable duration ratio model; Proceedings of International Conference on Acoustics, Speech, and Signal Processing 2006; 341-344.

[20] Sagisaka Y., Sato H.; Accentuation rules in Japanese TTS conversion; Rev. Elect. Commun. Lab. 1984; 32:188-199.

[21] Low P.H., Vaseghi S., Application of microprosody models in TTS synthesis; Proceedings of International Conference on Spoken Language Processing 2002; 2413-2416.

[22] Hifny Y., Rashwan M., Duration modeling for Arabic TTS synthesis; Proceedings of International Conference on Spoken Language Processing 2002; 1773-1776.

[23] El-Imam Y.A., Synthesis of the intonation of neutrally spoken modern standard Arabic speech; Signal Processing 2008; 88:2206-2221.

[24] Frid J., Prediction of intonation patterns of accented words in a corpus of read Swedish news through pitch contour stylization; Proceedings of European Conference on Speech Communication and Technology 2001; 915-918.

[25] Lobanov B., Tsirulnik L., Zhadinets D., Piorkovska B., Rafalko J., Szpilevsky E., Language-specific application of intonation contours in Russian and Polish multilingual TTS synthesis; Proceedings of International Conference on Speech and Computer 2005; 317-320.

[26] Kaiki N., Mimura K., Sagisaka Y., Statistical modeling of segmental duration and power control for Japanese; Proceedings of European Conference on Speech Communication and Technology 1991; 625-628.

[27] Fukuda T., Komori Y., Aso T., Ohora Y., A study of pitch pattern generation using HMM-based statistical information; Proceedings of International Conference on Spoken Language Processing 1994; 723-726.

[28] Fujio S., Sagisaka Y., Higuchi N.Z., Stochastic modeling of pause insertion using context-free grammar; Proceedings of International Conference on Acoustics, Speech, and Signal Processing 1995; 604-607.

[29] Taylor P., Black A.W., Assigning phrase breaks from part-of-speech sequences; Computer Speech and Language 1998; 12:99-117.

[30] Bulyko I., Ostendorf M., Joint prosody prediction and unit selection for concatenative speech synthesis; Proceedings of International Conference on Acoustics, Speech, and Signal Processing 2001; 781-784.

[31] Adell J., Agüero P.D., Bonafonte A., Database pruning for unsupervised building of text-to-speech voices; Proceedings of International Conference on Acoustics, Speech, and Signal Processing 2006; 889-892.

[32] Sakai S., Shu H., A probabilistic approach to unit selection for corpus-based speech synthesis; Proceedings of International Speech Communication Association Conference 2005; 81-84.

[33] Scordilis M.S., Gowdy J.N., Neural network-based generation of fundamental frequency contours; Proceedings of International Conference on Acoustics, Speech, and Signal Processing 1989; 219-222.

[34] Taylor P., Using neural networks to locate pitch accents; Proceedings of European Conference on Speech Communication and Technology 1995; 1345-1348.

[35] Riedi M., A neural-network-based model of segmental duration for speech synthesis; Proceedings of European Conference on Speech Communication and Technology 1995; 599-602.

[36] Callan D., Tajima K., Callan A., Akahane-Yamada R., Masaki S., Neural processes underlying perceptual learning of difficult second language phonetic contrast; Proceedings of European Conference on Speech Communication and Technology 2001; 145-148.

[37] Teixeira J.P., Freitas D.; Use of phoneme dedicated artificial neural networks to predict segmental durations; Proceedings of International Conference on Speech and Computer 2005; 679-682.

[38] Espinosa H.P., Reyes Garcia C.A., Genetic algorithms for the selection of cellular automata rules applied to intonation modeling in text to speech synthesis; Proceedings of International Conference on Speech and Computer 2005; 671-674.

[39] Wutiwiwatchai C., Furui S., Thai speech processing technology: a review; Speech Communication 2007; 49:8-27.

[40] Chen K., Hasegawa-Johnson M., Cohen A., An automatic prosody labeling system using ANN-based syntactic-prosodic model and GMM-based acoustic-prosodic model; Proceedings of International Conference on Acoustics, Speech, and Signal Processing 2004; 509-512.

[41] Ma X., Zhang W., Zhu W., Shi Q., Jin L., Probability based prosody model for unit selection; Proceedings of International Conference on Acoustics, Speech, and Signal Processing 2004; 649-652.

[42] Sun X., Applebaum T.H., Intonational phrase break prediction using decision tree and N-gram model; Proceedings of International Conference on Acoustics, Speech, and Signal Processing 2000; 1281-1284.

[43] Jiang D., Shi Q., Meng F., Shuang Z., Ma X., Lin Y., Overview of the IBM Mandarin text-to-speech system; Proceedings of Technology and Corpora Workshop on Speech-to-Speech Translation 2006; 181-185.

[44] Bahl L.R., de Souza P.V., Gopalakrishnan P.S., Nahamoo D., Decision tree for phonological rules in continuous speech; Proceedings of International Conference on Acoustics, Speech, and Signal Processing 1991; 185-188.

[45] Sheikhan M.; RNN-based prosodic information synthesizer for Farsi TTS; Second Irano-Armenian Workshop on Neural Networks, Dec. 1999.

[46] Chen S.H., Wang Y.R., Vector quantization of pitch information in Mandarin speech; IEEE Trans. Commun. 1990; 38:1317-1320.

[47] Childers D.G., Time modification of speech, theory: speech analysis, segmentation and labeling, In Speech Processing and Synthesis Toolboxes; John Wiley & Sons, Inc.; 2000:330-359.

[۴۸] فرخی علی، قائم‌مقامی شاهرخ، طبیانی محمود، شیخان منصور؛ تقطیع خودکار سیگنال گفتار در سطوح مختلف جمله به منظور استخراج اطلاعات نوا؛ مجموعهٔ مقالات دهمین کنفرانس مهندسي برق ایران (گرایش کامپیوتر) ۱۳۸۱؛ ۴۲۴-۴۳۱.

[49] Ghahramani Z., Jordan M.I., Supervised learning from incomplete data via an EM approach, In Advances in Neural Information Processing Systems; MIT Press; 1994:120-127.

[50] Weiss B., Prosodic elements of a political speech and its effects on listeners; Proceedings of International Conference on Speech and Computer 2005; 127-130.

[51] Moulines E., Emerard F., Larreur D., Le Saint Milon J.L., Le Fancheur L., Marty F., Charpentier F., Sorin C., A real-time French TTS system generating high quality synthetic speech; Proceedings of International Conference on Acoustics, Speech, and Signal Processing 1990; 309-312.

[52] Bulut M., Narayanan S.S., Syrdal A.K., Expressive speech synthesis using a concatenative synthesizer; Proceedings of International Conference on Spoken Language Processing 2002; 1265-1268.

[53] Erro D., Moreno A., A pitch-asynchronous simple method for speech synthesis by diphone concatenation using the deterministic plus stochastic model; Proceedings of International Conference on Speech and Computer 2005; 321-324.

[54] Toda T., Kawai H., Tsuzaki M., Optimizing sub-cost functions for segment selection based on perceptual evaluations in concatenative speech synthesis; Proceedings of International Conference on Acoustics, Speech, and Signal Processing 2004; 657-660.

[55] Nukaga N., Komoshida R., Nagamatsu K., Kitahara Y., Scalable implementation of unit selection based text-to-speech system for embedded solutions; Proceedings of International Conference on Acoustics, Speech, and Signal Processing 2006; 849-852.

[56] Storm V., Clark R., King S., Expressive prosody for unit-selection speech synthesis; Proceedings of International Speech Communication Association Conference 2006; 1296-1299.

[57] Conkie A., Syrdal A.K., Expanding phonetic coverage in unit selection synthesis through unit substitution from a donar voice; Proceedings of International Speech Communication Association Conference 2006; 1754-1757.

[58] Mohammadi M., Sheikhan M., TTS in broadcasting; Proc. International Conference BroadcastAsia 2000; 35-37.

[59] O'Brien D., Monaghan A., Concatenative synthesis based on a harmonic model; IEEE Trans. Speech Audio Processing 2001; 9:11-20.

[60] Zeljkovic I., Stylianou Y., Single complex sinusoid and ARHE model based pitch extractors; Proceedings of European Conference on Speech Communication and Technology 1999; 2813-2816.

[61] Stylianou Y., Removing linear phase mismatches in concatenative speech synthesis; IEEE Trans. Speech Audio Processing 2001; 9:232-239.

[62] Quatieri T.F., McAulay R.J., Shape invariant time-scale and pitch modification of speech; IEEE Trans. Signal Processing 1992; 40:497-510.

[63] Dutiot T., Stylianou Y., Text-to-speech synthesis, In: Oxford Handbook of Computational Linguistics; Oxford University Press; 2003:323-338.

[64] O'Brien D., Monaghan A., Shape invariant time-scale modification of speech using a harmonic model; Proceedings of International Conference on Acoustics, Speech, and Signal Processing 1999; 381-384.

[65] ITU-T Recommendation P. 85: A method for subjective performance assessment of the quality of speech output devices; International Telecommunications Union Publication; 1994.

[66] Alvarez Y.V., Huckvale M., The reliability of the ITU-T P. 85 standard for the evaluation of TTS systems; Proceedings of International Conference on Spoken Language Processing 2002; 329-332.