مدل کردن کشش زماني واج برای سیستم تبدیل متن به گفتار فارسی به کمک شبکه عصبي

مجید نم نبات، عباس کوچاری، سعید رحیمی، بهرنگ قاسمی زاده، مرتضی سالاریان[1]
{namnabat,kochari, qasemizadeh,rahimi, salaraian}@ digitalclone.net

چکیده

در این مقاله از یک شبکه عصبی چند لایه پرسپترون برای مدل کردن کشش زماني استفاده شده است. بدین منظور ابتدا 60 دقیقه گفتار برچسب گذاری شد و از آن برای داده های آموزشی و آزمایشي استفاده گردید. ویژگیهاي مورد استفاده برای پیشگویي شبکه عصبی فقط شامل ویژگیهای واجي و هجایي می باشند و از اطلاعات نحوي که مربوط به نقش کلمه در جمله می باشد استفاده نشده است. بعد از انتخاب ویژگیهای اولیه، بازنگری بر روی ویژگیها صورت گرفت و تعداد ویژگیها کاهش داده شد. سپس از دو روش تغییر مقیاس تقسیم بر ماکزیمم و تبدیل z-score برای نرمال کردن مقدار کشش زماني استفاده گردید. براي ارزیابي از معیار میزان همبستگي میان کشش زماني پیشگویي شده و مقدار واقعی استفاده شده است که برای روش تقسیم بر ماکزیمم و تبدیل z-score به ترتیب برابر 77.82% و 80.39% بدست آمد.

کلمات کلیدي: شبکه های عصبی چند لایه پرسپترون[2]، تبدیل z-score، کشش زمانی، تبدیل متن به گفتار.

1- مقدمه

یکي از بخشهاي مهم سیستمهاي تبدیل متن به گفتار[3]، تخمین نواي متن ورودي مي باشد. اجراي اصلي نواي گفتار عبارتند از میزان کشش واجها، منحني پیچ جملات و شدت صدا و عناصر فرعي آن ریتم صحبت، سرعت بیان و تن صدا مـي باشـند. بطـور کلـي پیشگویي پارامترهاي نواي گفتار بسیار دشوار و پیچیده مي باشد. این پیچیدگي اولاً ناشي از نامشخص بـودن تـاثیرات متقابـل اجـزاي مختلف نوا بر یکدیگر و دوماً ناشي از وقایع نادر[4] مي باشد. بررسیهای تجربی نشان دهنـده ارتبـاط پیچیـده و نامشـخص میـان اجـزای مختلف نوا برای انتقال مفاهیم گفتار می باشند. بواقع انتقال مفاهیم گفتار توسط گویندگان با ترکیب تمامی پارامترهای مختلف نوا بـه شنوندگان صورت می گیرد. بطور مثال تمامی اجزای نوای گفتار در تعیین اجزای با اهمیت گفتار نسبت به بخشهای دیگر نقش دارنـد. مسئله دیگر وجود وقایع نادر براي مدل کردن نواي گفتار مي باشد. این وقایع هر چند به ندرت اتفاق مي افتند ولـي مجمـوع فرکـانس وقوع تمامی آنها به اندازه اي مي باشد که همیشه حداقل یک واقعه نادر در هر جمله روي مي دهد، لذا این وقایع نبایـد نادیـده گرفتـه شوند[3]. در سیستمهاي تبدیل متن به گفتار کنوني عمدتاً از مدل کردن اجـزاي پیچیـده نـوا همچـون ریـتم و سـرعت بیـان گفتـار صرفنظر مي شود و علاوه بر این ارتباط اجزاي اصلي نوا بصورت مستقل در نظر گرفته مي شود. عموماً در ایـن سیسـتمها ابتـدا میـزان

[1] اعضای تیم سیستم تبدیل متن به گفتار کلونایزر

[2] Multi-Layer Perceptron Neural Networks (MLP)

[3] Text To Speech (TTS)

[4] Rare Events

کشش زماني واجها تعيين و سپس با استفاده از ميزان کشش زماني واجها، منحني پيچ جملات تعيين مي شود. پيشگويي دقيق ميزان کشش زمانی، علاوه بر افزايش ميزان طبيعی بودن خروجی سيستم در افزايش کارايي مدلهای تخمين پيچ نيز به دليل وابستگی آن به کشش زمانی نقش مهمی دارد.

روشهای تخمين کشش زمانی را می توان به دو دسته عمده روشهای قانون-گرا[5] و روشهای آماری تقسيم بندی نمود. کلات اولين سيستم تخمين کشش زمانی را بصورت قانون-گرا در سيستم تبديل متن به گفتار MITalk ارائه نمود[1,2]. اين سيستم برای بسياری از زبانهای ديگر پياده سازی و مورد استفاده قرار گرفت. امروزه با بزرگتر شدن دادگانهای گفتار، روشهای آماری مورد توجه قرار گرفته اند. روشهای آماری را می توان به دو دسته مدلهای پارامتری و غير پارامتری تقسيم بندی نمود. در مدلهای پارامتری ساختار پردازش ويژگيهای ورودی مشخص است. بطور مثال می توان به مدل مجموعی از حاصلضربها[6] اشاره نمود[13]. در اين روش تاثير ويژگيهای مختلف در ميزان کشش زمانی بصورت يک حاصل جمع از تاثير ويژگيهای مختلف در نظر گرفته می شود. تاثير هر ويژگي نيز بصورت حاصلضرب تاثير اين ويژگی نسبت به ويژگيهای ديگر مدل می شود[10]. عمده مدلهای غيرپارامتری شبکه های عصبی و درختهای تصميم گيری[10] و منحني های رگرسيون متعدد تطبيقی[7][6] می باشند که برای پيشگويي کشش زمانی مورد استفاده قرار گرفته اند. بررسيها نشان می دهد که شبکه های عصبی و منحنی های رگرسيون متعدد تطبيقی کارايي بهتری نسبت به درختهای تصميم گيری دارند[7,8]. کمپل واحد تخمين کشش را هجا بدليل پايدارتر بودن آن نسبت به واج در نظر گرفت. وی برای محاسبه ميزان کشش زمانی واجها بر اساس کشش زمانی هجا از مفهوم z-score استفاده نمود[4]. استفاده از z-score برای مدل کردن کشش زمانی امکان مدل کردن سرعت بيان گفتار را نيز براحتی فراهم می نمايد[5]. روش MARS روش جديدی می باشد که برای پيشگويي کشش زمانی کارايي مناسبی در حدود 90% برای زبان آلمانی داشته است. اين روش برای مدل کردن داده ها از مجموع يکسری منحنی های رگرسيون استفاده می کند. آموزش اين روش نيازمند حجم محاسباتی بالا و زمانبر می باشد[7]. چانگ برای مدل کردن کشش زمانی از درختهای تصميم گيری و رگرسيون استفاده کرده است. وی بردار ويژگی مربوط به حروف صدادار را از صامت جدا ساخته و برای هر يک درخت تصميم گيری جداگانه ای ساخته است[9].

در اين پژوهش از شبکه های عصبی چند لايه پرسپترون برای تخمين کشش زمانی واجها استفاده شده است و دو روش مختلف تغيير مقياس براي تخمين کشش زماني بررسی شده است. در بخش 2 اين مقاله دادگان مورد استفاده، در بخش 3 ويژگيهای انتخاب شده شرح داده می شوند. در بخش 4 مدل شبکه عصبی شرح داده می شوند. بخش 5 به ارزيابی سيستم اختصاص يافته است و در انتها بخش 6 به نتيجه گيری و ارائه پيشنهادات می پردازد.

2- دادگان مورد استفاده

برای تهيه داده های مورد نياز بمنظور آموزش و تست سيستم از دادگان فارس دات بزرگ استفاده شده است. در اين دادگان 100 گويشور با تنوع سن، جنس، ميزان تحصيلات مختلف و همچنين با لهجه های متعدد صحبت کرده اند. هر گويشور بطور متوسط حدود 4000 کلمه از متون منتخب از روزنامه های کثيرالانتشار در داخل کشور را با توجه به تنوع موضوعی بطور رسمی خوانده است. ضبط صدا در يک محيط آرام و با فرکانس نمونه برداری 22050 هرتز انجام شده است. برای استخراج داده ها از 60 دقيقه گفتار يک گويشور مرد از اين دادگان استفاده شده است. سيگنال صحبت اين گوينده بطور دستی در سطح واج به دقت تقطيع شده است. برای استخراج ويژگيهای ورودی لازم است که متن اين گفتار برچسب گذاری شود. برای اينکار ابتدا کلمات سازنده متن استخراج و با رشته آوانگاری صحبت آن تراز شده اند. سپس اين متن و کلمات آن در سطح واج، هجا، کلمه برچسب گذاری شده اند. لازم به ذکر است که اطلاعات نحوی مورد استفاده قرار نگرفته اند و تخمين کشش زماني مستقل از نقش کلمات در جمله و براساس اطلاعات واجی و هجاهاي کلمه صورت گرفته است.

---

[5] Rule-Based
[6] Sum-of-Products
[7] Multivariate Adaptive Regression Splines (MARS)

3- ویژگیها

برای ارزیابی کارایي ویژگیهای مختلف، 35 ویژگي از دادگان متنی استخراج گردید. سپس تاثیر هر یک از این ویژگیها در کارایي مدل شبکه عصبی مورد ارزیابی قرار می گیرد تا در نهایت از یک زیر مجموعه بهینه از این ویژگیها برای آموزش هر یک از مدلها استفاده شود. این ویژگیها عبارتند از:

- شماره شناسه واج مورد نظر و واجهای همسایه آن: در زبان فارسی 30 واج وجود دارد که با یک شناسه عددی می توان آنها را از هم تفکیک کرد. طول همسایگی برای هر واج در کلمه، دو در نظر گرفته شده است. این بدین معنی است که علاوه بر شناسه واج مورد نظر، از شناسه های دو واج بعدی و قبلی آن نیز بعنوان ویژگي استفاده می شود.

- ویژگیهای آوایي واج مورد نظر و واجهای همسایه: این ویژگیها عبارتند از کلاس آوایي هر واج، محل تولید آن، بیواک، واکدار یا صدادار بودن آن. برای صدادارها، ویژگي کلاس آوایي متمایز کننده واجهای صدادار کشیده و کوتاه از یکدیگر در نظر گرفته شده است. علاوه بر این ویژگي محل تولید برای صدادارها آنها را نظر مشابهت میان ویژگیهای ثانویه آنها از یکدیگر متمایز می نماید.

- موقعیت واج مورد نظر در کلمه و هجا و انتهای هجا: ویژگی موقعیت واج در کلمه برای متمایز ساختن واجهای اول، دوم، آخر و یکی مانده به آخر و بقیه موقعیتها در کلمه از یکدیگر به کار می رود. هر هجا را می توان به سه بخش ابتدایي[8]، هسته[9] و انتهایي[10] تقسیم بندی نمود. ویژگی موقعیت واج در هجا مشخص کننده بخشی از هجا می باشد که واج مورد نظر به آن تعلق دارد. با توجه به ساختار هجابندی زبان فارسی، دو بخش ابتدایي و هسته همواره شامل یک واج می باشند ولی بخش انتهایي می تواند حداقل صفر و حداکثر دو واج داشته باشد. در صورتیکه واج مورد نظر متعلق به بخش انتهایي باشد، ویژگی موقعیت در انتهای هجا، جایگاه آن را در این بخش مشخص می نماید.

- ویژگیهای آوایي هجای شامل واج مورد نظر و هجاهای همسایه: در زبان فارسی نوع هسته هجا و طول بخش انتهایي آن مشخص کننده ویژگیهای هجا می باشند. برای تعیین هسته هجا دو ویژگی نوع و شماره شناسه هسته در نظر گرفته شده است. نوع هسته کشیده یا کوتاه بودن هسته و شماره شناسه متمایز کننده واجهای صدادار با یک نوع می باشد. این ویژگي علاوه بر هجای شامل واج مورد نظر، برای هجای قبلی و بعدی آن نیز استخراج شده است.

- تعداد هجاها و موقعیت هجای شامل واج در کلمه: برای موقعیت هجا در کلمه سه حالت ابتدا، انتها و وسط کلمه بودن در نظر گرفته شده است. علاوه بر این حالتها، بعضی کلمات می توانند تک هجایي باشند که برای این هجاها نیز یک حالت در نظر گرفته شده است. برای مشخص کردن تعداد هجاهای کلمه شامل واج مورد نظر نیز پنج حالت در نظر گرفته شده است. کلمات با تعداد هجای کمتر از چهارتا با حالتهای جداگانه و کلمات با بیش از چهار هجا نیز با یک حالت مشخص می شوند.

3-1- بازنگري ویژگیها

بعد از آموزش با داده هاي آموزشي و تست کردن، نتایج حاکي از این نکته داشت که استفاده از بعضي از ویژگیها باعث کندي سیستم و حتي کاهش کارایي مي گردد. در این مرحله براي داشتن ویژگیهاي مناسب که از آنها بتوان بهترین مدل را آموزش داد، با تغییر ویژگیهاي آموزشي بهترین آنها که نتایج بهتري و هزینه کمتري دارند انتخاب گردید. ویژگیهاي انتخابي 15 ویژگي مي باشند که بعضي از آنها از ترکیب ویژگیها بدست آمده اند و بیانگر ویژگیهاي دیگر با صرف هزینه کمتر مي باشند. بررسیها نشان می دهند که در میان واجهای همسایه، خصوصیات واج دو تا قبل و همچنین ویژگیهای هجای بعدی دارای تاثیر بسیار کم و قابل اغماض در کارایي مدلها می باشند.

---

[8] Onset
[9] Nuclear
[10] Coda

جدول 1- ویژگیهای استفاده شده برای شبکه عصبي

سطح واجشناسی	ویژگي	تعداد نورون
واج	کلاس آوايي واج قبلي	8
	محل توليد واج فعلي	12
	ترکيب بيواک، واکدار يا صدادار بودن واج فعلي	3
	کلاس آوايي واج قبلي	8
	محل توليد واج قبلي	12
	ترکيب بيواک، واکدار يا صدادار بودن واج قبلي	3
	کلاس آوايي واج بعدي	8
	محل توليد واج بعدی	12
	ترکيب بيواک، واکدار يا صدادار بودن واج بعدي	3
	کلاس آوايي واج دو تا بعد	8
	محل توليد واج دو تا بعد	12
	ترکيب بيواک، واکدار يا صدادار بودن واج دو تا بعد	3
	مکان واج در انتها	2
هجا	نوع هسته هجا	2
	نوع انتهاي هجا	3

4- شبکه عصبی MLP

در ابتدا از یک شبکه عصبی چند لایه پرسپترون برای تعیین کشش زمانی واجها استفاده شد. این شبکه عصبي یک شبکه پیشرو با چهار لایه مي باشد که داراي یک لایه ورودي، یک لایه خروجي و دو لایه مخفي است. دو لایه مخفی به ترتیب دارای 30 و 6 نورون در نظر گرفته شده است. در لایه ورودی 104 نورون و در لایه خروجی برای تعیین میزان کشش واجها یک نورون در نظر گرفته شده است. تابع فعالیت لایه های ورودی و میانی، تانژانت هایپربولیک و لایه خروجی به صورت خطی انتخاب شده است. برای مقدار دهی اولیه وزنها و مقادیر بایاس لایه اول از الگوریتم گوین-ویدرو[11] استفاده شده است[11]. در این الگوریتم از میانگین داده های آموزشی بعنوان مقادیر اولیه وزنها و بایاس استفاده میشود. در جدول 1 ویژگیهای بهینه که برای آموزش شبکه انتخاب شده است همراه با تعداد نورونهای اختصاص یافته به هر ویژگی نشان داده شده است. تمامی این ویژگیها گسسته می باشند و برای ارائه این ویژگیها به شبکه های عصبی بهتر است برای هر مقدار از هر ویژگی یک عضو از بردار ورودی اختصاص داده شود. در اینحالت با توجه به مقدار هر ویژگی، یک عضو از اعضای اختصاص یافته از بردار ورودی به آن ویژگی یک و بقیه صفر در نظر گرفته میشوند.

بردارهای ورودی و خروجی شبکه های عصبی بهتر است که در محدود [0,1] و یا [-1,1] تغییر نمایند. برای تغییر مقیاس مقدار کشش واجها می توان به دو روش عمل کرد. در روش اول می توان بطور ساده مقدار کشش واجها را بر ماکزیمم مقدار کشش تقسیم نمود تا محدوده تغییر آنها [0,1] شود. در روش دوم از تبدیل z-score برای تغییر محدوده استفاده می شود. برای انجام این تبدیل لازم است که لگاریتم میانگین (μ_i) و لگاریتم انحراف معیار (σ_i)کشش برای هر واج محاسبه گردد. سپس مقدار z-score کشش زمانی (dur) بصورت زیر بدست می آید:

$$Dur_{z-score(\log)} = (\log(dur) - \mu_{\log(i)}) / \sigma_{\log(i)} \qquad (1)$$

[11] Nguyen-Widrow

در فرمول فوق i بیانگر یک واج است. در صورتیکه توزیع لگاریتم کشش زمانی نرمال باشد با استفاده از تبدیل فوق در حدود 98% داده ها در محدوده [-2,2] قرار می گیرند. استفاده از لگاریتم باعث می شود که توزیع نرمال بهتری برای کشش زمانی بدست آید. استفاده از تبدیل z-score برای تخمین میزان کشش زمانی، امکان مدل کردن سرعت بیان و حالتهای بیان مختلف را با تغییر میانگین و انحراف معیار هر واج را براحتی فراهم می سازد. با استفاده از این تبدیل تاثیر میزان کشش ذاتی هر واج حذف و امکان تخمین میزان کشش واجها بهتر فراهم می شود.

برای آموزش شبکه از الگوریتم انتشارخطا به عقب جهنده[12] استفاده شده است. الگوریتم انتشار خطا به عقب جهنده علاوه بر کارآیي مناسب در حد الگوریتم لونبرگ-مارکوارت[13] نسبت به آن حافظه کمتری مصرف میکند و برای شبکه های با داده های آموزشی زیاد مناسب می باشد، در صورتیکه برای لایه های میانی شبکه MLP تابع فعالیت سیگموئید انتخاب شود. در روند آموزش آن با الگوریتم انتشار خطا به عقب به دلیل کوچک بودن شیب این تابع برای ورودیهای بزرگ، دامنه گرادیان بسیار کوچک میشود و مقدار تغییر وزنها و بایاس در هر بار آموزش بسیار کوچک می شود. در این الگوریتم دامنه گرادیان یک مقدار ثابت انتخاب میشود و برای تعیین آن از شیب تابع فعالیت استفاده نمی شود[12]. برای جلوگیری از یادگیری بیش از حد داده های آموزشی[14] توسط شبکه ها از 15% داده های آموزشی برای ارزیابی[15] سیستم استفاده شده است.

5- پیاده سازی و نتایج

برای پیاده سازی شبکه های عصبی از نرم افزار مطلب استفاده شده است. معیار ارزیابي تعیین دقت میزان کشش زماني تخمین زده شده، مقدار همبستگي[16] مي باشد. فرمول محاسبه این مقدار در زیر دیده مي شود.

$$Correlation = \frac{\sum_i (x-\mu_x)(y-\mu_y)}{\sqrt{\sum_i (x-\mu_x)^2}\sqrt{\sum_i (y-\mu_y)^2}} \qquad (2)$$

در جدول 2 مقدار همبستگی بدست آمده براساس روش تغییر مقیاس و z-score دیده می شود. ویژگیهاي استفاده شده در هر دو روش تغییر مقیاس، ویژگیهای جدول 1 می باشد.

جدول 2- مقدار همبستگی دو روش مختلف برای پیشگویي کشش زماني

روش	تقسیم بر ماکزیمم	z-score لگاریتمی
کارایي برای مجموعه آموزشی	%78.31	%78.99
کارایي برای مجموعه تست	%77.82	%80.39

6- نتیجه گیری و ارائه پیشنهادات

در این مقاله از شبکه های عصبی چندلایه برای پیشگویي کشش زمانی که از پارامترهای اصلی نوای گفتار است، استفاده شده است. برای اینکار از ویژگیهای استخراجی از دادگان فارس دات برای آموزش و آزمایش مدلها استفاده گردید. انتخاب بهینه ویژگیها نقش بسیار مهمي در تخمین مناسب کشش زماني دارد. در اینجا مقدار کشش زمانی بصورت کلي و با نرمال کردن کشش زماني و هم با استفاده از معیار z-score مورد استفاده قرار گرفت که نتایج حاصل نشان از برتري استفاده از معیار z-score نسبت به روش دیگر

[12] Resilient Backpropagation
[13] Levenberg-Marquardt
[14] Overfitting
[15] Validation
[16] Correlation

دارد. علت این برتري در اینست که در معیار z-score از میانگین و واریانس هر واج برای تخمین کشش زماني استفاده می شود که از انحراف بیش از حد مقدار پیشگویي شده جلوگیري می کند. پیشنهادي که برای ادامه کار مطرح است دسته بندی آواها به دو دسته صدار دار و بی صدا و تشکیل دو شبکه عصبي مجزا برای هر دسته می باشد. علاوه بر این می توان از دیگر روشهای یادگیری ماشین همچون درختهای تصمیم گیری و یا روش MARS برای تخمین کشش زمانی استفاده کرد. علاوه بر این اضافه کردن ویژگیهای نحوی به ویژگیهای انتخابی به احتمال زیاد باعث افزایش کارایي خواهد شد.

مراجع

[1] D. H. Klatt, “Review of Text-to-Speech Conversion for English”, J. Acoustical Society of America, vol. 82, no. 3, pp. 737-793, 1987

[2] D. H. Klatt, “Linguistic Uses of Segmental Duration in English: Acoustic and perceptual Evidence”, Journal of Acoustic Society of America, vol. 59, pp. 1209–1221, 1976.

[3] B. Moebius, J. Van Santen, “Modeling Segmental Duration in German Text-to-Speech Synthesis”, Spoken Language Conference, Vol. 4, no. 2395-2398, 1996.

[4] W. N. Campbell, “Predicting Segmental Durations for Accommodation within a Syllable-Level Timing Framework”, EuroSpeech, 1993.

[5] R. Cordoba, J. A. Vallejo, J. M. Montero, J. Arriola-Gutierrez, M. A. Lopez, J. M. Pardo, “Selection of the Most Significant Parameters for Duration Modelling in a Spanish Text-to-Speech System Using Neural Networks”, Computer Speech and Language, Vol. 16, pp. 183-203, 2002.

[6] J. H. Friedman, “Multivariate Adaptive Regression Splines”, The Annals of Statistics, vol. 19, no. 1, pp. 1-141, 1991.

[7] M. Riedi, “Modeling Segmental Duration with Multivariate Adaptive Regression Splines”, Eurospeech, vol. 5, pp. 2627-2630, 1997.

[8] M. Riedi, "A Neural-Network Based Model of Segmental Durational for Speech Synthesis, Eurospeech, vol. 1, pp. 599-602, ESCA, 1995.

[9]H. Chung, “Duration Models and the Perceptual Evaluation of Spoken Korean”, Speech Prosody, pp. 219-222, 2002.

[10] Lee, S. and Y.H. Oh, “Tree-based modeling of prosodic phrasing and segmental duration for Korean TTS systems”, Speech Communication, vol. 28, pp. 283-300, 1999.

[11] H. Demuth, M. Beale, “Neural Network Toolbox for Use with Matlab”, Users Guide Version 3.0, 1998.

[12] M. Riedmiller, H. Braun, “A Direct Adaptive Method for Faster Back-Propagation Learning: The RPROP algorithm”, IEEE International Conference on Neural Networks, San Francisco, CA, 589-591, 1993.

[13] V. Santen, “Assignment of Segmental Duration in Text-To-Speech Synthesis”, Computer, Speech and Language, vol. 8, pp. 95-128, 1994.