

# مدلسازی تنوعات گویش در سیستم بازشناسی گفتار پیوسته

## با استفاده از شبکه عصبی

بهرام وزیرنژاد<sup>۱</sup>، فرشاد الماس گنج<sup>۲</sup>

دانشگاه صنعتی امیرکبیر، دانشکده مهندسی پزشکی  
{bvazirnezhad, almas}@cic.aut.ac.ir

### چکیده

در این مقاله در صدد هستیم به منظور تولید گونه‌های تلفظی کلمات برای ایجاد واژگان حاوی گونه‌های تلفظی کلمات، در سیستم بازشناسی گفتار پیوسته فارسی، از شبکه عصبی استفاده نماییم. این شبکه عصبی بعد از آموزش قادر است تلفظ‌های احتمالی را با در اختیار داشتن نسخه مرجع واجی کلمه پیشگویی نماید. در این روش شبکه عصبی می‌تواند برای هر نسخه واجی مرجع تعدادی گونه‌های تلفظی تولید نماید. در این مقاله نحوه آموزش شبکه عصبی و بکارگیری آن در جهت تولید گونه‌های تلفظی کلمات به همراه احتمالات آنها برای جایگزینی در واژگان سیستم، تشریح می‌گردد. نتایج آزمایشات بیانگر تطبیق بهتر گونه‌های تلفظی تولید شده با ورودی‌ها است که افزایش درصد صحت بازشناسی کلمات با استفاده از واژگان حاوی گونه‌های تلفظی تولید شده نسبت به حالت استفاده از واژگان بدون گونه تلفظی را بدنبال خواهد داشت.

**کلید واژه:** مدلسازی تنوعات تلفظی، سیستم بازشناسی گفتار پیوسته، شبکه عصبی

### ۱- مقدمه

تولید واژگانی که حاوی گونه‌های تلفظی کلمات باشد یکی از نکات اساسی در تلاش برای بهبود عملکرد سیستم بازشناسی گفتار است. یکی از اولین تلاشها برای این منظور استفاده از قواعد واجی بود که در TBM انجام پذیرفت [۱]. هنوز هم تولید واژگان حاوی تلفظ‌های کلمات یکی از راهکارهای مؤثر برای بهبود عملکرد سیستمهای بازشناسی گفتار در زمینه‌های واژگان بزرگ محسوب می‌شود. تولید گونه‌های تلفظی کلمات بصورت دستی و یا با استفاده از قواعد تلفظی احتیاج به زمان زیادی دارد و کار پیچیده‌ای است. بنابراین تحقیقات زیادی در راستای تولید خودکار واژگان حاوی گونه‌های تلفظی کلمات صورت گرفته است. در سالهای آغازین دهه ۱۹۹۰، با ایجاد پایگاه دادگانی نظیر TIMIT که به‌صورت فونتیکی واج‌نویسی شده بودند. محققانی نظیر راندولف [۲]، ریلی [۳]، ووترز و استولک [۴] تلاشهای زیادی برای مدلسازی تلفظ نمودند [۲]. اگر چه در نتیجه این تلاشها امکان تولید خودکار قواعد تلفظی که تغییرات واجی را توصیف می‌نمودند، بوجود آمد ولی هنوز به واج‌نویسی فونتیکی به‌صورت دستی احتیاج بود. در تلاشهای بعدی با استفاده از یک سیستم بازشناس واج که می‌تواند نسخه‌های فونتیکی را ایجاد نماید، امکان استفاده از دادگان بزرگ مهیا شد. چنین تلاشهایی توسط اشמיד در سال ۱۹۹۳، اسلوبودا [۵]، امایی [۶] و هامفریز [۷] انجام گرفت. بسیاری از محققان نیز از درختهای تصمیم برای مدلسازی تلفظ بهره جستند (راندولف [۸]، ریلی [۹]، فوسلر [۱۰]، وین‌تراب [۱۱]، هامفریز [۱۲] و بیرن [۱۳]).

<sup>۱</sup> دانشجوی دکتری مهندسی پزشکی، گرایش بیوالکتریک

<sup>۲</sup> استادیار دانشکده مهندسی پزشکی، گرایش بیوالکتریک

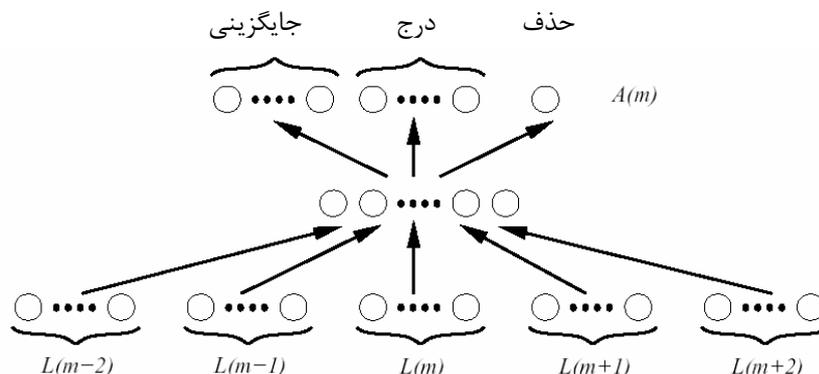
در سالهای اخیر، با تلاش برای ایجاد سیستمهای با واژگان بزرگ با کاربرد بازشناسی گفتار محاوره‌ای مانند گفتار موجود در دادگان Switch board، مدلسازی تلفظ نکته بسیار مهمی است. چرا که تلفظ کلمات در گفتار محاوره‌ای دارای تنوعات بسیار بیشتری نسبت به گفتار روخوانی است [۴]. در این مقاله برای تولید خودکار واژگان تلفظی از دادگان بزرگ گفتاری فارسی‌دات استفاده شد. و به عنوان سیستم بازشناس واج از شبکه عصبی بازشناسی واج سیستم شنوا برای تولید نسخه‌های فونتیک استفاده گردید. روش مدلسازی تنوعات، بهره‌گیری از شبکه عصبی برای یادگیری تفاوت‌های واجی بین نسخه واجی فونتیک و نسخه واجی مرجع بوده است. بعد از آموزش، این شبکه قادر است با در اختیار داشتن نسخه واجی مرجع به تولید گونه‌های تلفظی کلمات بپردازد. در تلاش‌های قبلی برای تولید گونه‌های تلفظی از روش استخراج قواعد تلفظی بهره جستیم [۵]. در این مقاله نحوه امتیازدهی برای محاسبه احتمالات تغییرات تلفظی و راهکار جدیدی برای مدلسازی تغییرات تلفظی در مرز کلمات مورد توجه قرار گرفته است. در بخش ۲ نحوه آموزش شبکه و تولید گونه‌ها توسط آن را مورد بررسی قرار می‌دهیم. در بخش ۳ نحوه استفاده از گونه‌های تلفظی را در بازشناسی گفتار پیوسته بیان می‌نماییم. بخش ۴ به ذکر آزمایشات طراحی شده و ارزیابی مدل تلفظی، اختصاص یافته است. در انتها طی بخش ۵ به بحث و نتیجه‌گیری در مورد تحقیق انجام گرفته پرداخته می‌شود.

## ۲- تولید خودکار واژگان حاوی گونه‌های تلفظی کلمات

در این بخش نحوه آموزش شبکه عصبی مورد استفاده و نیز نحوه تولید گونه‌های تلفظی کلمات توسط این شبکه بررسی می‌گردد. دو اصطلاح مورد استفاده در بخش‌های بعدی را به صورت زیر تعریف می‌نماییم.  
**تلفظ مرجع:** دنباله واجی مرجع در کلمه است. فرض می‌شود کلمه در روخوانی به این نحو تلفظ می‌گردد. و تغییرات تلفظی ناشی از تنوع گوینده و اثرات هم‌تولیدی آواها در گفتار محاوره‌ای را در نظر نمی‌گیریم.  
**تلفظ بیان شده:** دنباله واجی که به صورت واقعی توسط گوینده بیان گردیده است. این دنباله واجی به دلایلی نظیر تنوعات بین گوینده‌ها و تنوع گویشها و اثرات هم‌تولیدی آواها از دنباله واجی مرجع کلمه متفاوت است.

### ۲-۱- شبکه تلفظی

برای پیشگویی تلفظ‌های بیان شده از تلفظ مرجع، از یک شبکه عصبی چند لایه پرسپترون مانند شکل ۱ استفاده نمودیم. در این تحقیق، برای هر واج موجود در تلفظ مرجع یا  $L(m)$  یک واج بیان شده یا  $A(m)$  بصورت متناظر، تولید می‌گردد. این کار با توجه به متن واجی یعنی ۴ واج اطراف  $L(m)$  شامل  $L(m+2), \dots, L(m-2)$  انجام می‌شود. در بخش ۲-۲ نحوه آموزش شبکه تلفظی را بیان می‌نماییم و در بخش ۲-۳ فرآیند تولید گونه‌های تلفظی کلمات با استفاده از این شبکه عصبی بررسی می‌گردد.



شکل ۱- ساختار شبکه عصبی مورد استفاده برای یادگیری تنوعات تلفظی

## ۲-۲- نحوه آموزش

### ۲-۲-۱- تهیه دادگان آموزش

دادگان مورد استفاده برای آموزش تنوعات تلفظی به شبکه عصبی، شامل ۹۰۰ فایل wave برچسب‌دار شده از فارس دات بزرگ<sup>۲</sup> بوده است. این مجموعه شامل محدوده وسیعی از گوینده‌ها شامل ۵۰ مرد و ۵۰ زن با سنین از ۱۱ تا ۵۵ ساله و با تحصیلات گوناگون بوده است. گوینده‌ها دارای لهجه‌های گوناگون شامل تهرانی، ترکی، اصفهانی، یزدی و ... بودند و متنهایی که توسط گوینده‌ها روخوانی می‌شد محدوده وسیعی از مقولات فرهنگی، سیاسی، اجتماعی و ... را دربرمی‌گرفت بنابراین سعی گردید تا مجموعه وسیعی از کلمات رایج در زبان فارسی مورد استفاده قرار گیرند. در نهایت حدود ۱۷۵۰۰۰ کلمه برای آموزش شبکه، مورد استفاده قرار گرفت. برای آموزش شبکه تلفظی، ابتدا باید دادگان آموزش آماده شود، که عبارت از مجموعه بزرگی زوج ورودی (تلفظ مرجع) و خروجی (تلفظ بیان شده) می‌باشد. داده لازم با تولید دنباله‌های واجی تلفظهای بیان شده از طریق بازشناسی آواها و همردیف سازی آن با دنباله واج‌های تلفظهای مرجع کلمات متناظر ایجاد می‌گردد. این فرآیند را می‌توان به مراحل زیر بیان نمود.

- ۱- بازشناسی آواهای موجود در دادگان گفتار پیوسته فارس دات با استفاده از سیستم بازشناسی واج صورت می‌گیرد. دنباله واج‌های بازشناسی شده همان دنباله واج‌ها در تلفظهای بیان شده را تشکیل می‌دهند.
- ۲- همردیف‌سازی دنباله‌های واجی تلفظهای مرجع با دنباله‌های واجی تلفظهای بیان شده با استفاده از یک الگوریتم همردیف‌سازی انجام می‌گیرد. برای مثال، اگر نتیجه بازشناسی واج برای تلفظ دنباله مرجع واجی /a r a y u r u/ به صورت دنباله واج /a w a u r i u/ باشد، نتیجه همردیف‌سازی این زوج دنباله بصورت زیر خواهد بود.

$a r a y u r u$	دنباله واجی مرجع
$a w a u r i u$	دنباله واجی بیان شده

در این دنباله واج  $r$  از دنباله مرجع به واج  $w$  در تلفظ تبدیل شده، و واج  $y$  حذف شده و واج  $i$  برای واج ششم درج گردیده است. بنابراین  $A(2) = w, L(2) = r, A(4) = \phi, L(4) = y, A(6) = \{r, i\}$  که  $r$  درج شده است. واجهایی که به همان صورت مرجع تلفظ شده‌اند نیز بصورت تبدیل واج استخراج شده‌اند. برای مثال  $a$  بصورت  $a$  تلفظ شده است. بازشناسی واج برای تمامی دادگان انجام گرفته و نتایج دنباله‌های بازشناسی شده با تلفظهای مرجع همردیف شده‌اند تا دادگان بصورت زوج ورودی - خروجی آماده گردد.

### ۲-۲-۲- ساختار شبکه تلفظی

برای آموزش شبکه تلفظی، متن واجی شامل پنج واج از تلفظ مرجع، یا  $L(m+2), \dots, L(m-2)$  بعنوان ورودی در نظر گرفته می‌شوند. و  $A(m)$  که با  $L(m)$  همردیف شده است خروجی شبکه خواهد بود. ۱۴۵ نرون (۲۹ واج زبان فارسی به ازای هر یک از ۵ واج ورودی) در لایه ورودی استفاده می‌شوند. در لایه خروجی یک نرون نمایانگر حذف واج، ۲۹ نرون به تعداد واجهای فارسی نمایانگر واج جایگزین و ۲۹ نرون دیگر نمایانگر واج درج شده در صورت وجود می‌باشند. بنابراین ۵۹ نرون در لایه خروجی وجود دارد. با توجه به مثال قبلی نماد حذف واج که به چهارمین واج از تلفظ مرجع مربوط شده است. بعنوان  $A(m)$  و دنباله /r a y u r/ به عنوان واج‌های متنی  $L(m+2), \dots, L(m-2)$  در نظر گرفته می‌شود. در اینجا در فاز آموزش خروجی مطلوب مربوط به حذف واج ۱ در نظر گرفته می‌شود. نرون مربوط به  $r$  در دسته  $L(m-2)$ ، نرون مربوط به  $a$  در دسته  $L(m-1)$ ، نرون مربوط به  $y$  در دسته  $L(m)$ ، نرون مربوط به  $u$  در دسته  $L(m+1)$  و نرون مربوط به  $r$  در دسته  $L(m+2)$  نیز با ورودی ۱ تغذیه می‌شوند. بقیه ورودی و خروجی‌ها ۰ در نظر گرفته می‌شوند.

### ۲-۳- فرآیند تولید گونه‌های تلفظی

فرض کنید می‌خواهیم بهترین گونه تلفظی را از خروجی‌های شبکه تلفظی پیدا نماییم. اگر دنباله واجی تلفظ مرجع کلمه  $W$  بصورت  $L = [L(1), \dots, L(|W|)]$  باشد. و  $|W|$  تعداد واج‌های موجود در تلفظ مرجع باشد ( $|W| \leq 5$ ). دنباله واجی تلفظ بیان شده بصورت زیر توسط شبکه پیشگویی می‌شود.

۱-  $i$  را مساوی ۳ قرار بده،  $A(1) = L(1)$  و  $A(2) = L(2)$ .

۲- برای ۵ واج متنی مربوط به واج  $i$  ام یا  $L(i+2), \dots, L(i-2)$ ، به نرون‌های ورودی مربوط به هر یک از واجها از ۱۴۵ نرون لایه ورودی شبکه عصبی (عدد یک) اعمال شود.

۳- اندیس نرون خروجی که بیشترین سطح فعالیت را دارد مشخص شود.

۳-۱- اگر نرون با بیشترین سطح فعالیت از نرونهای جایگزین واج بود.  $A(i)$  را معادل واج مربوط به این نرون قرار بده.

۳-۲- اگر نرون با بیشترین سطح فعالیت از نرونهای درج واج بود. از مجموعه نرونهای خروجی مربوط به جایگزینی واج نرون دوم را با بیشترین سطح فعالیت را پیدا کن. و  $A(i)$  را معادل مجموعه دو واجی مربوط به این دو نرون به ترتیب استخراج شده، قرار بده.

۳-۳- اگر نرون با بیشترین سطح فعالیت در لایه خروجی همان نرون نمایانگر حذف واج باشد،  $A(i)$  را معادل حذف واج قرار بده.

۴- شمارنده  $i$  را یک واحد افزایش بده.

۵- مراحل ۲ تا ۴ را تکرار کن تا زمانیکه  $i = |W| - 1$  باشد.

۶-  $A(|W|) = L(|W|)$  و  $A(|W| - 1) = L(|W| - 1)$ .

با این روش می‌توان چندین گونه تلفظی، برای هر کلمه پیدا نمود. گونه‌های با احتمال بیشتر برای جایگیری در واژگان انتخاب می‌گردند. برای مثال خروجی‌های شبکه برای دنباله مرجع /arigato/ در جدول ۱ آمده است.

جدول ۱- نمونه ورودی-خروجی شبکه برای دنباله واجی مرجع /arigato/ بیانگر حذف است.

واج (خروجی-خروجی هنجارسازی شده)			ورودی
۳	۲	۱	
o (۰/۱۱-۰/۱)	u (۰/۲۲-۰/۲)	i (۱-۰/۹)	ariga
b (۰/۲۵-۰/۱)	g (۰/۷۵-۰/۳)	$\phi$ (۱-۰/۴)	rigat
o (۰/۲۵-۰/۲)	e (۰/۵-۰/۴)	a (۱-۰/۸)	igato
k (۰/۴-۰/۲)	d (۰/۶-۰/۳)	t (۱-۰/۵)	gato

### ۳- استفاده از گونه‌های تلفظی در بازشناسی گفتار پیوسته

سیستم‌های بازشناسی گفتار، کلمه بازشناسی شده را با توجه به رابطه  $\hat{w} = \arg \max_{w \in W} p(w|o)$  انتخاب می‌نمایند. در این مقاله، با توجه به اینکه گونه‌های تلفظی کلمات را به همراه احتمالات آنها مدل نموده‌ایم، رابطه فوق را بصورت زیر تعمیم می‌دهیم.

$$\hat{w} = \arg \max_{w \in W} \sum_{A \in W} p(A, w|o) \quad (1)$$

که در این رابطه مجموع روی  $A$  یا گونه تلفظی کلمه  $w$ ، می‌باشد. با توجه به قانون بیز، سمت راست رابطه (۱) می‌تواند بصورت زیر نوشته شود.

$$\arg \max_{w \in W} \sum_{A \in W} p(o|A, w)P(w)P(A|w) \quad (2)$$

$p(o | A, w)$  احتمال شرطی بروز ویژگیهای صوتی  $O$  به شرط بیان دنباله واجی تلفظ  $A$  از کلمه  $w$  است. این احتمال می‌تواند از طریق یک مدل صوتی محاسبه گردد.  $p(w)$  از طریق مدل زبانی قابل دستیابی است و می‌توان آنرا از طریق مدل  $n$ -gram محاسبه نمود.  $p(A | w)$  را مدل تلفظی می‌نامیم. در این مقاله شبکه عصبی بعنوان مدل تلفظی بکار گرفته شده است. تنوعات ناشی از گوینده‌های مختلف و شرایط متفاوت می‌تواند از طریق ذکر چندین گونه تلفظی در واژگان مدل گردد. اما با این فرض که یک گوینده خاص در شرایط مشخص کلمات را به نحو ثابتی ادا می‌نماید، می‌توان مجموع را از رابطه قبلی حذف نمود و به رابطه ۳ دست یافت.

$$\arg \max_{w \in W, A \in \mathcal{A}} P(o | A, w) P(w)^\alpha P(A | w)^\beta \quad (3)$$

$\alpha$  و  $\beta$  ضرایب وزندهی برای مدل زبانی و مدل تلفظی می‌باشند [۶].

#### ۴- ارزیابی مدل تلفظی

برای ارزیابی مدل تلفظی، گونه‌های تلفظی تعدادی از کلمات به همراه احتمال آنها را تولید نمودیم. و نحوه تطابق گونه‌های تولید شده را با گونه‌هایی که در عمل بصورت واقعی ایجاد شده‌اند را با استفاده از یک الگوریتم برنامه پویا بدست آوردیم. بعد از هم‌ردیف‌سازی گونه‌های تولیدی مدل با گونه‌های واقعی به ازای هر عدم تطابق بین واجها هزینه ۱ در نظر گرفته شد. سپس میزان تطابق گونه‌های تولیدی مدل را با میزان تطابق دنباله مرجع با گونه‌های واقعی مقایسه نمودیم. نتایج بیانگر آن است که گونه‌های تولید شده توسط مدل بسیار بیشتر از دنباله مرجع با گونه‌های واقعی تطابق دارند. لذا می‌توان نتیجه گرفت آوردن گونه‌های تلفظی کلمات در واژگان سیستم می‌تواند درصد صحت بازشناسی کلمات را افزایش دهد.

برای مثال در مورد کلمه "اطلاعات" با دنباله واجی / *et te l a t a t a* /، ۷ گونه از مدل بدست آمد و ۳ گونه واقعی نیز در دادگان دیده شده است. هر یک از ۷ گونه مدل با هر یک از ۳ گونه واقعی هم‌ردیف‌سازی می‌شوند. و سپس هزینه هم‌ردیف‌سازی‌ها مجموع گرفته می‌شود. در انتها کمیت به‌دست آمده بر  $7 \times 3$  تعداد هم‌ردیف‌سازی‌ها تقسیم گردید.

$$\text{هزینه عدم تطابق گونه‌های تولید کلمه اطلاعات با گونه‌های واقعی} = \frac{\sum_{t=1}^{7 \times 3} \cos t [DP(A_m, A_r)] \times P(A_m | W)}{7 \times 3} = 1.8 \quad (4)$$

که  $A_m$  بیانگر گونه‌های تولیدی مدل و  $A_r$  بیانگر گونه‌های واقعی است.

میانگین هزینه‌های هم‌ردیف‌سازی دنباله مرجع با گونه‌های واقعی نیز به طریق زیر بدست آمد.

$$\text{هزینه عدم تطابق گونه دنباله واجی مرجع با گونه‌های واقعی} = \frac{\sum_{t=1}^3 \cos t (DP(L, A_r))}{3} = 2.8 \quad (5)$$

با توجه به آزمایشات انجام گرفته گونه‌های تولید شده کلمات توسط مدل بسیار بهتر با گونه‌های واقعی همان کلمات تطابق دارند. لذا می‌توان به راحتی استنتاج نمود، ذکر گونه‌های تلفظی کلمات در واژگان می‌تواند به افزایش درصد صحت بازشناسی کلمات کمک شایانی نماید.

#### ۵- بحث و نتیجه‌گیری

در این مقاله به ارائه روشی برای تولید خودکار گونه‌های تلفظی کلمات، با استفاده از شبکه عصبی پرداختیم. شبکه عصبی مورد استفاده با استفاده از دادگان فارس دات آموزش داده شد. نحوه تولید چندین گونه برای هر کلمه را به همراه احتمالات آنها ذکر نمودیم. نتایج آزمایشات بیانگر آن است که گونه‌های تولید شده توسط مدل با گونه‌های واقعی تطابق خوبی دارند. و نسبت به حالتی که فقط گونه مرجع کلمه در واژگان سیستم استفاده شود، نتایج بهتری قابل انتظار است.

آزمایشات نشان می‌دهد، میانگین هزینه هم‌ردیف‌سازی گونه‌های تولید مدل با گونه‌های واقعی بسیار کمتر از میانگین هزینه هم‌ردیف‌سازی دنباله واجی مرجع کلمه با گونه‌های واقعی می‌باشد. با توجه به این آزمایشات می‌توان انتظار داشت، درصد صحت بازشناسی کلمات، با استفاده از واژگان حاوی گونه‌های تلفظی تولیدی مدل افزایش خواهد داشت. البته باید در انتخاب تعداد گونه‌های کلمات دقت زیادی به خرج داد. چراکه کارایی سیستم می‌تواند با افزایش بیش از حد گونه‌های تلفظی کاهش یابد. این امر بدلیل ایجاد سردرگمی سیستم، در انتخاب می‌باشد. ضمن آنکه هزینه محاسباتی نیز با افزایش تعداد گونه‌ها افزایش پیدا می‌کند. باید توجه داشت، تخصیص احتمالات مناسب به گونه‌ها، نقش ویژه‌ای در کاهش نرخ خطای بازشناسی دارد. مزیت اساسی این روش نسبت به تولید دستی گونه‌های تلفظی، وجود احتمالات گونه‌های تلفظی بر اساس دادگان است. در روش دستی مبتنی بر اطلاعات زبانشناسی، این امکان بصورت دقیق وجود ندارد.

## سپاسگزاری

از مسئولین محترم پژوهشکده پردازش هوشمند علائم، به ویژه جناب آقای دکتر جواد شیخ‌زادگان به دلیل در اختیار گذاشتن امکانات و دادگان فارس‌دات بسیار سپاسگزاریم.

## مراجع

- [1] L. Bahl, J. Baker, P. Cohen, F. Jelinek, B. Lewis, R. Mercer, "Recognition of a Continuously Read Natural Corpus", *Proc. ICASSP 78*, pp.422-424, 1978 .
- [2] C. Wooters, A. Stolcke, "Multiple Pronunciation Lexical Modeling in a Speaker Independent Speech Understanding System", *Proc. ICSLP-94*, pp.1363-1366, 1994 .
- [3] T. Imai, A. Ando, E. Miyasaka, "A New Method for Automatic Generation of Speaker-dependent Phonological Rules", *Proc. ICASSP 95*, pp.864-867, 1995 .
- [4] T. Fukada, Y. Sagisaka, "Automatic Generation of a Pronunciation Dictionary based on a Pronunciation Network", *Proc. EUROSPEECH-97*, pp.2471-2474, 1997 .
- [5] ب. وزیرنژاد، ف. الماس‌گنج، "مدلسازی گونه‌های تلفظی کلمات در سیستم بازشناسی اتوماتیک گفتار پیوسته فارسی" در مجموعه مقالات اولین کنفرانس بین‌المللی فن‌آوری اطلاعات و دانش، تهران، ۱۳۸۲.
- [6] T. Fukada, T. Yoshimura, Y. Sagisaka, "Automatic Generation of Multiple Pronunciations based on Neural Networks", *J. of Speech Communication*, Vol. 27, pp.63-73, 1999 .