بهبود بازشناسی گفتار پیوسته فارسی با استفاده از خصوصیات پروزودیک گفتار در سطح واج‌ها

فرشاد الماس گنج
استادیار
دانشکده مهندسی پزشکی، دانشگاه صنعتی امیرکبیر

کارولوکس
استاد
دانشکده فنی ومهندسی، دانشگاه تهران

محمود بی جن خان
استادیار
دانشکده ادبیات و علوم انسانی، دانشگاه تهران

سیّدمحمّدرضا هاشمی گلپایگانی
استاد
دانشکده مهندسی پزشکی، دانشگاه صنعتی امیرکبیر

سیّدعلی سیّدصالحی
استادیار
دانشکده مهندسی پزشکی، دانشگاه صنعتی امیرکبیر

چکیده

در بازشناسی خودکار واج های گفتار پیوسته فارسی مشکلاتی مشاهده می شوند که بسیاری از آنها از ماهیت سیگنال گفتار ناشی می شوند و وابستگی زیادی به روش بازشناسی واج ندارند. در این گزارش به شرح روش هایی پرداخته ایم که برای رفع تعدادی از این مشکلات فراهم آمده اند. ویژگی هایی که برای این منظور به کار رفته اند از نوع ویژگی های پروزودیک گفتار در سطح واج ها می باشند. افزودن این روش ها به روش عادی ما در بازشناسی واج ها که از ترکیب شبکه های عصبی و قواعد آوایی زبان فارسی شکل گرفته است، باعث گردید که افزایشی بیش از ۱/۵ درصد در صحت بازشناسی واج ها برای ۱۴۰ جمله آزمایشی که از دادگان «فارس دات» برداشته شده اند، داشته باشیم و در مجموع به درصد صحت بازشناسی واج ۷۴/۸ درصد برای این ۱۴۰ جمله برسیم.

Better Phoneme Recognition by Using Phoneme's Prosodic Features

F. Almasganj
Assistant Professor
Amirkabir University

C. Lucus
Professor
Tehran University

S. M. R. Hashemi Golpaygani
Professor
Amirkabir University

M. Bijankhan
Assistant Professor
Tehran University

S. A. S. Salehi
Assistant Professor
Amirkabir University

Abstract

Many of the errors that occur in automatic phoneme recognition of continuous speech recongnition are arisen from the complicated nature of speech signal such that the known phoneme recognition approaches can not handle them. Phoneme prosodic features can be used to alter some of these errors. In this paper, we report 1.5% decreasing of phoneme recognition error rate by using some prosodic features on the level of phonemes, such as energy, aspiration and lips protrusion. This approach has evaluated on a set of 140 test sentences that are taken from "FARSDAT" database, yeilding 74.8% phoneme recongnition accuracy rate.

کلمات کلیدی

پروزودی - بازشناسی واج - بازشناسی گفتار.

مقدمه

در بازشناسی واج های گفتار پیوسته فارسی مشکلاتی رخ می دهند که روش های عمده بازشناسی در حل برخی از آنها دچار مشکل می شوند. این نوع مشکلات که از ماهیت سیگنال گفتار ناشی می شوند، خوب است به طور جداگانه و به عنوان مسائل خاص بازشناسی واج ها مورد مطالعه قرار گیرند و راه حل مخصوص هر کدام مهیا گردد.

آنچه که در این گزارش به آن پرداخته ایم، معرفی چند مورد مهم از این نوع مشکلات و ارائه راه حل های مناسب برای آنهاست. موارد مورد بحث با استفاده از ویژگی هایی که از انواع ویژگی های پروزودیک در سطح واج ها و خصوصیات اسپکتروگرام[1] سیگنال تشکیل شده اند، مورد بررسی قرار گرفته و راه حل هایی جهت بازشناسی دقیق تر آنها عنوان گردیده اند.

ویژگی های پروزودیک استفاده شده از دو نوع ویژگی های پروزودیک کلاسیک و غیر کلاسیک می باشند. منحنی فرکانس پایه شدت سیگنال صوتی در کنار اثراتی که مربوط به دو عامل سرعت و ریتم هستند، در مجموع به عنوان خصوصیات پروزودیک کلاسیک گفتار [۱] شناخته می شوند. غیر از این تعریف، نوع گسترده تری از خصوصیات پروزودیک در سطح واج ها تعریف می شوند که آنها را با عنوان ویژگی های پروزودیک غیر کلاسیک در سطح واج ها نام می بریم. این تعریف فراگیر پروزودی در سطح واج ها [۲] به صورت زیر عنوان می شود:

پروزودی در سطح واج ها مجموعه ای از وضعیت های دستگاه تولید گفتار است که می تواند به صورت لایه ای مستقل بر روی لایه واج ها اثر گذارد.

هر واج از تعدادی وضعیت تولیدی دستگاه گفتار تشکیل شده است که اینها می توانند بر روی واج های مجاور خود اثر گذارند. باتوجه به این تعریف از پروزودی، هر کدام از این وضعیت های تولیدی که بر روی واج های قبلی یا بعدی خود اثر می گذارند، به عنوان یک عامل پروزودیک تلقی می شوند. به این ترتیب لایه پروزودی شامل ویژگی هایی است که از لحاظ بازشناسی واج ها حشو هستند. برای مثال واج /p/ دارای ویژگی هایی است که در هر حالتی از تولید /p/ موجود هستند و نشان می دهند که واج تولیدی /p/ است. ولی ویژگی های دیگری هستند که بنا بر موقعیت قرار گرفتن /p/ در گفتار تولید شده و یا حذف می گردند، مثل دمش[3]. این ویژگی ها تأثیری در /p/ بودن این واج ندارند، بلکه حاوی اطلاعات اضافه ای می باشند. این نوع ویژگی ها را می توان در لایه پروزودی در سطح واج ها قرار داد.

در این گزارش روش هایی ارائه شده اند که از دو نوع ویژگی دمش و گردی لب[4] در کنار ویژگی پروزودیک طول زمانی و خصوصیات اسپکتروگرام سیگنال گفتار در جهت افزایش درصد صحت بازشناسی واج ها استفاده می کنند.

برای ارزیابی روش های عنوان شده از یک داده گان تست ۱۴۰ جمله ای از ۱۴ گوینده مختلف، برداشته شده از داده گان «فارس دات» [۳] استفاده نموده ایم. داده های یادگیری مورد لزوم برای بخش هایی که نیاز به اطلاعات یادگیری دارند، نیز از «فارس دات» برداشته شده اند. برای اندازه گیری درصد صحت بازشناسی واجها از برنامه استاندارد "NIST" استفاده نموده ایم که مخفف عبارت "National Institute of Standard and Technology" می باشد.

۲ - روش کار

۲ - ۱ - جداسازی واکه های به هم چسبیده

در فرایند بازشناسی واج های گفتار پیوسته فارسی به مواردی برخورد می کنیم که نشانگر واکه های طولانی می باشند. تعدادی از این موارد مربوط به حالتی می شوند که همخوان بین دو واکه (معمولاً /r/یا/h/?/) به علت کوتاه یا ناسوده تلفظ شدن، حذف شده است و دو واکه به صورت یک واکه طولانی بازشناسی شده اند. این موارد را می توان با مقایسه کشش واکه بازشناسی شده با میانگین کشش آن واکه مشخص نمود.

واحدی که برای اندازه گیری سرعت تولید گفتار پیوسته بکار برده ایم، برحسب تعداد هجاهای تولیدی در یک ثانیه می باشد. متوسط سرعت تولید جملات در دادگان «فارس دات» برابر ۴/۶۳۷ هجا بر ثانیه می باشد، [۴] که آن را جهت نرمالیزه کردن سرعت جملات مختلف به کار می بریم. برای مثال اگر جمله ای سریعتر از این حد بیان شده باشد، کشش واج های آن را به نسبتی افزایش می دهیم که سرعت جمله مورد نظر به میانگین ذکر شده برسد.

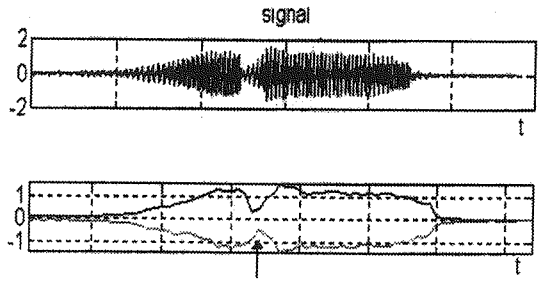
بعد از نرمالیزه کردن سرعت تولید جملات مختلف «فارس دات» ، میانگین کشش واکه های زبان فارسی به صورت جدول (۱) به دست می‌آیند [۴].

جدول (۱) میانگین کشش واکه های زبان فارسی در گفتار پیوسته (نرمالیزه شده)

نوع واکه	میانگین کشش واکه (msec)
A	۱۴۱/۳۶
a	۱۱۹/۷۷
e	۸۰/۹۴
o	۱۰۱/۶۰
u	۱۱۶/۸۰
i	۱۰۶/۱۰

برای محاسبه میانگین کشش واکه های مختلف از دادگان «فارس دات» استفاده نموده‌ایم. باید توجه کرد که در مقایسه کشش واکه های بازشناسی شده با این مقادیر متوسط، سرعت تولید گفتار نقش مهمی را ایفا می‌کند و کشش واکه ها قبل از مقایسه با متوسط موجود باید ابتدا نسبت به سرعت تولید گفتار نرمالیزه شوند و سپس مقایسه صورت گیرد.

اگر کشش واکه بازشناسی شده بیش از ۱/۵ برابر میانگین کشش آن واکه بود، امکان اینکه این واکه از دو واکه تشکیل شده باشد، افزایش می‌یابد. در چنین مواردی به بررسی تغییرات لحظه‌ای دامنه سیگنال صوتی این واکه طولانی می‌پردازیم. همین عمل را در مواقعی که روش بازشناسی واج های مورد استفاده ما امکان حضور یک همخوان در وسط واکه بازشناسی شده را با عدم قطعیت اعلام داشته است، می‌توانیم مورد استفاده قرار دهیم. شکل ۱ سیگنال صوتی کلمه «ورود» را به همراه مرزهای پایین و بالای انرژی آن و نشانه‌ای که حد فاصل بین واکه /0/و /u/ است را نشان می‌دهد. مطابق این شکل در حدفاصل بین دو واکه /o/ و /u/ یک کاهش قابل ملاحظه در دامنه سیگنال دیده می‌شود که مشخص کننده مرز بین دو واکه است. در مواردی که این کاهش دامنه به بیش از ۳۰٪ حداکثر دامنه (این آستانه به طور تجربی تعیین شده است) برسد، واکه طولانی بازشناسی شده را به دو وا که مجزا تبدیل می‌کنیم و مرز اینها را در محل کاهش دامنه سیگنال قرار می‌دهیم.



شکل (۱) سیگنال صوتی کلمه «ورود» به همراه مرزهای بالا و پایین دامنه آن.

۲-۲-دمش

یکی از تفاوت های مهم واج های /k/,/t/,/p/ به ترتیب با واج های /g/, /d/,/b/ در وجود ویژگی «دمش» در گروه اول است. وقوع دمش در گروه اول که همخوان هایی بی واک[۵] هستند، یک عامل متمایز کننده این دو گروه همخوان ازیکدیگر است و می‌تواند در فرایند اصلاح واج ها (در مواقعی که بین همخوانی از گروه اول یا دوم، تردید وجود دارد) مؤثر واقع شود.

برای بازشناسی «دمش» در این واج ها از یک شبکه عصبی چهار لایه پس انتشار خطا [۵] استفاده کرده‌ایم. باتوجه به اینکه «دمش» در بخش انتهایی همخوان های انفجاری بی واک رخ می‌دهد، بازنمایی های مربوط به ۷ فریم اخر واج های مورد نظر را در ورودی شبکه عصبی قرار داده‌ایم.

هر فریم با ۱۸ پارامتر معرفی می‌شود. این پارامترها توسط مجموعه ای از فیلترها که با مقیاس «بارک»[۶] تنظیم شده‌اند، تولید می‌گردند [۶]. برای استخراج این بازنمایی، از فریم هایی ۲۳ میلی ثانیه ای سیگنال بعد از اعمال پنجره همینگ[۷] تبدیل فوریه گسسته می‌گیریم. سپس مجموعه ای از ۱۸ فیلتر که در مقیاس «بارک» [۷] مراکز آنها در $Z_k=k$ قرار گرفته است به تبدیل فوریه هر فریم اعمال می نماییم تا ۱۸ پارامتر مورد نظر به دست آیند. تبدیل محور فرکانس از هرتز «بارک» به کمک رابطه زیر انجام می‌شود:

$$Z = 6\ ln\ [f/600 + sqrt\ ((f/600)^2 + 1)] \qquad (۱)$$

شکل فیلترها در حوزه فرکانس با مقیاس «بارک» به صورت مجذور پنجره «هنینگ»[۸] در فواصل مشخص

شده زیر است:

در فاصله $Z_k + 1 > Z > Z_k - 1$ $C_k = (0.5 + 0.5 \cos(\Pi (Z - Z_k)))^2$ (۲)

و در خارج این فاصله $C_k = 0$

مشخصات شبکه عصبی استفاده شده در جدول (۱) آمده است:

جدول (۱) مشخصات شبکه عصبی بازشناسی کننده «دمش».

لایه خروجی	لایه پنهان دوم	لایه پنهان اول	لایه ورودی
۲	۷۲	۲۲۴	۷×۱۹=۱۳۳

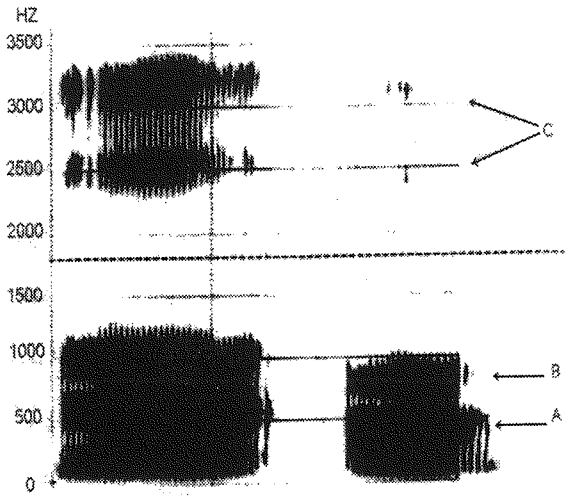
فعال شدن یکی از نرون های خروجی شبکه به معنی وجود «دمش» و فعال شدن دیگری به معنی عدم وجود «دمش» است. در مواردی که هر دو نرون فعال شوند، تصمیم نهایی براساس نرونی گرفته می شود که فعال تر است و خروجی بالاتری دارد.

اطلاعات تعلیم شبکه از دادگان «فارس دات» انتخاب شده اند و شامل تعداد زیادی واج انفجاری از نوع «واکدار» و «بی واک» از گویندگان مختلف می شوند.

بعد از اتمام مراحل مختلف یادگیری شبکه عصبی، آن را بر روی واج های انفجاری «داده آزمون» که از ۱۴۰ جمله از دادگان «فارس دات» تشکیل می شود، آزمایش نمودیم. درصد بازشناسی «دمش» برابر با ۹۳/۵ درصد به دست آمد.

۲-۳- گردی لب ها

در موضوع گردی «لب ها» مورد اشتباه شدن واکه های /o/ و /A/ را که مورد بسیار مهمی است و در حین بازشناسی واج ها به تعداد قابل توجهی رخ می دهد، انتخاب نمودیم. از تفاوت های مهم این دو واکه، گردی بیشتر لب در واکه /o/ را می توان مورد توجه قرار داد. ویژگی گردی لب در واکه /o/ بسیار قویتر از واکه /A/ ظاهر می شود. این ویژگی باعث کم شدن فاصله فورمنت های اول و دوم می شود [۷]. تفاوت دیگری هم بین این واکه ها وجود دارد و آن ویژگی «ارتفاع زبان» است. ارتفاع زبان و فورمنت اول نسبت به هم حالت معکوس دارند [۸]. بنابر این فورمنت اول در واکه /o/ که ارتفاع زبان در آن بالاتر است، پایین ترمی باشد. شکل ۲ اسپکتروگرام دو واکه /o/ و /A/ را نشان می دهد.



شکل (۲) از سمت چپ به ترتیب اسپکتروگرام دو واکه /o/ و /A/ ـ نشانه های A و B به ترتیب فورمنت های اول و دوم واکه ها را نشان می دهند.

همانگونه که انتظار می رود فورمنت های اول و دوم واکه /o/ پایین تر از /A/ هستند. نکته دیگری که در این شکل دیده می شود، تضعیف فرکانس های بالاتر در /o/ می باشد که اثر دیگر گردی لب است (در محل نشانه C).

باتوجه به تفاوت هایی که در اسپکتروگرام این دو واکه دیده می شوند، ویژگی هایی را از روی طیف انرژی سیگنال صوتی دو واکه انتخاب نموده ایم که باتوجه به آنها می توان تعیین نمود، با کدامیک از این واکه ها مواجه هستیم. این ویژگی ها در جدول ۲ آورده شده اند.

جدول (۲) مشخصات ویژگی هایی که برای جداسازی واکه های /o/ و /A/ به کار رفته اند.

ویژگی اول	«انرژی بالای ۱۱۰۰ هرتز» بر «انرژی زیر ۱۱۰۰ هرتز»
ویژگی دوم	«انرژی زیر ۵۰۰ هرتز» بر «انرژی بالای ۵۰۰ هرتز»
ویژگی سوم	«انرژی بین ۹۰۰ و ۱۳۲۰هرتز» بر «انرژی زیر ۴۵۰ هرتز»

متوسط و انحراف معیار این نسبت ها را برای این دو واکه در ۱۴۰ جمله «داده تست» محاسبه نمودیم. جدول ۳ این مقادیر را نشان می دهد.

جدول (۳) میانگین و انحراف معیار سه ویژگی از طیف انرژی واکه های /o/ و /A/

نوع ویژگی	نوع واکه	متوسط	انحراف معیار
ویژگی اول	A	۲۰۵	۱۴۴
ویژگی اول	O	۸۰	۲۴۹
ویژگی دوم	O	۳/۹	۲/۴
ویژگی دوم	A	۰/۵	۱/۲
ویژگی سوم	A	۸۵۷	۵۴۶
ویژگی سوم	O	۱۰۰	۲۷۶

در تعریف ویژگی های اول تا سوم به دنبال این بوده ایم که ویژگی هایی را بیابیم که بتوانند حداکثر فاصله را بین واکه های /o/ و /A/ ظاهر نمایند. بعد از یافتن میانگین و انحراف معیار این ویژگی ها برای واکه های /o/ و /A/ از تابع توزیع نرمال [۹] برای ساختن تابع امکان آنها سود می جوییم:

$$m(x) = e^{-(x - x_m)^2 / 2\sigma^2}$$

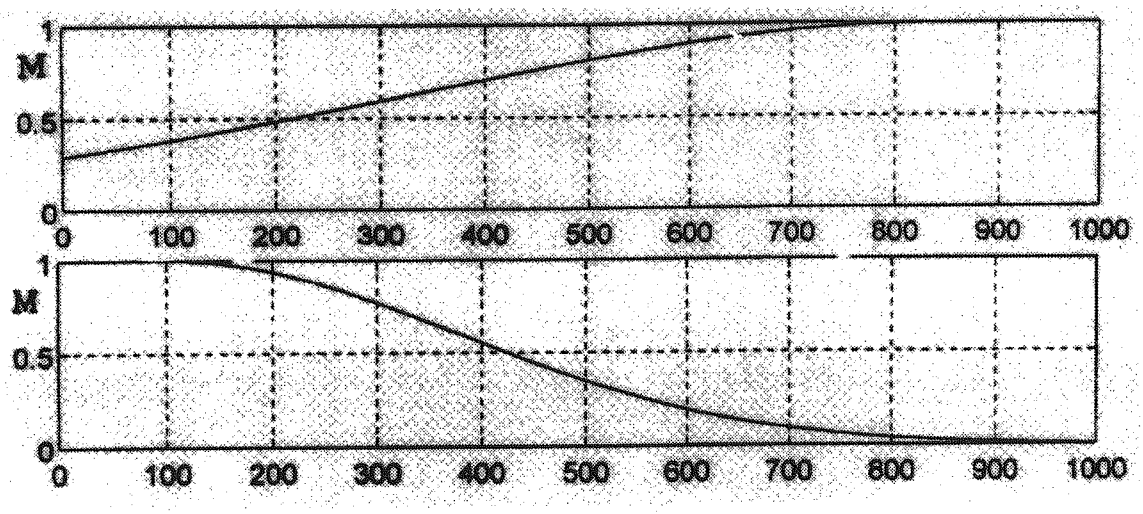
نمونه ای از توابع امکان /o/ و /A/ برای ویژگی سوم در شکل ۳ دیده می شوند. همانگونه که در شکل ۳ دیده می شود، امکان واکه /A/ بودن را برای مقادیر بالاتر از ۸۵۷، و امکان واکه /o/ بودن را برای مقادیر کمتر از ۱۰۰، برابر ۱ فرض کرده ایم.

بخش باقیمانده تابع امکان دو واکه را با استفاده از رابطه توزیع نرمال محاسبه نموده ایم. در محاسبه توابع امکان /o/ و /A/ برای ویژگی های دیگر هم از همین روش پیروی شده است. برای بازشناسی /o/ و /A/، ابتدا ویژگی های اول تا سوم را برای سیگنال واکه محاسبه می کنیم و سپس امکان /A/ یا /o/ بودن را با قرار دادن مقادیر این ویژگی ها در توابع امکان مربوطه، برای هر ویژگی به طور جداگانه محاسبه می نماییم. سپس به ترتیب وزن های ۰/۵، ۱ و ۱ را به مقادیر امکان محاسبه شده بر اساس ویژگی های اول تا سوم اعمال می کنیم و نتایج حاصل شده را با یکدیگر جمع می کنیم. تصمیم گیری نهایی در مورد اینکه این واکه a/ یا/A/ است، باتوجه به مجموع مقادیر امکان محاسبه شده برای آنها گرفته می شود.

در بازشناسی /A/ و /b/ در ۱۴۰ جمله «داده تست»، برداشته شده از دادگان «فارس دات» ۵۹ مورد اشتباه بین این واکه ها مشاهده می گردید که با کمک گرفتن از روش فوق به ۲۶ مورد کاهش یافت.

۳ـ نتیجه گیری

در این مقاله به روش هایی اشاره شد که با استفاده از ویژگی های پروزودیک و خصوصیات اسپکتروگرام سیگنال گفتار در سطح واج ها به اصلاح درصد صحت بازشناسی آنها می پرداختند. کارایی این روش ها با



شکل (۳) توابع امکان واکه های /o/ و /A/ باتوجه به ویژگی سوم، به ترتیب از بالا به پایین.

افـزایش ۱/۵ درصـدی صـحت بازشناسی واج ها و رساندن آن به مقدار ۷۴/۸ درصد که مقدار قابل توجهی است، نشـان مـی‌دهد کـه ویژگی‌های پروزودیک باید در بازشناسی گفتار فارسی مورد توجه بیشتری قرار گیرند وبا توسـعـه این رویکرد و اسـتـفـاده از ویژگی‌های پروزودیک بیشتر در سطح واج ها می‌توان به نتایج قابل توجـهی در زمینه اصـلاح روش‌های بازشناسی واج ها رسـیـد. به خـصـوص زمـانی کـه روش‌های کـلاسـیک بازشناسی گفتار به اشباع می رسند، این خصوصیات حاوی اطلاعـات فراوانی هستند که با اسـتـفـاده از آنها می‌توان بر بسیاری از مشکلات بازشناسی گفتار غلبه نمود.

زیرنویس‌ها

1- Spectrogram
2- Rhythm
3- Aspiration
4- Protrusion
5- Unvoiced
6- Bark
7- Hamming
8- Hanning

مراجع

[۶] سیدصالحی، س، ع ، «بازشناخت گفتار پیوسته فارسی با استفاده از مدل عملکردی مغز انسان در درک گفتار» رساله دکتری، دانشگاه تربیت مدرس، ۱۳۷۴.

[7] Hermansky, H., "Perceptual Linear Predictive (PLP) Analysis of Speech", Journal of Acoustical Society of America, 87 94)m pp. 1738-1752, 1990.

[8] Beddor, P.S., "Predicting the Structure of Phonological Systems", Journal of Phonetica, 48, pp. 83-107, 1991.

[9] Papoulis, A., "Probability, Random Variables and Stochastic Processes", Mc Graw-Hill Inc., 1991.

[1] Crystal, D. "Encyclopedia of Language and Linguistic", London, pp. 169-173, 1992.

[2] Lass, R. "Phonology and Introduction to Basic Concepts", Cambridge Universiy Press, London, 1983.

[3] Bijankhan, M., "FARSDAT-The Speech Data base of FARSI Spoken Language", SST'94, pp. 836-831, 1994.

[۴] الماس گنج، ف . «تجزیه و تحلیل ساختار گفته های زبان فـارسی با اسـتـفـاده از اطلاعـات پروزودیک سـیگنال گـفـتـار»، رسـاله دکـتـری، دانشگاه تربیت مدرس، ۱۳۷۷.

[5] Lippman, R. P. "An Introduction to Computing with Neural Nets", IEEE ASSP Mag., Appril, pp. 4-22, 1987.