علوم و مهندسی کامپیوتر
نشریه علمی پژوهشی انجمن کامپیوتر ایران
مجلد ۳، شماره ۳ (الف)، پائیز ۱۳۸۴
صفحات ۱۳-۲۲
مقاله عادی

بهبود بازشناسی گفتار با استفاده از تلفیق دانش واژگانی با اطلاعات صوتی

محمدرضا یزدچی[۱] سیدعلی سیدصالحی[۲] فرشاد الماس‌گنج[۳]

دانشکده مهندسی پزشکی، دانشگاه صنعتی امیرکبیر، تهران، ایران

چکیده

توانایی‌های بالای انسان در ادراک گفتار ما را به بهره‌گیری هرچه بیشتر از نحوه عمکرد مغز انسان در درک گفتار تشویق می‌نماید. در مقاله حاضر با استفاده از روش‌های معکوس‌سازی شبکه‌های عصبی و شبکه‌های استخراج کننده مؤلفه‌های اساسی غیرخطی مدل‌های واژگانی ایجاد می‌شود. استفاده از این مدل‌ها صحت بازشناسی واج مدل صوتی مرجع را در کلمات مجزا حداکثر تا ۸۱٪ افزایش می‌دهد.تلفیق نتایج به‌دست آمده از مدل واژگانی با دادگان صوتی، صحت بازشناسی واج را مجدداً افزایش می‌دهد.به این منظور و با استفاده از روش‌های معکوس‌سازی شبکه‌های عصبی نتایج مدل واژگانی با دادگان صوتی تلفیق می‌شود. اصلاح پارامترهای بازنمایی بر اساس نتایج حاصل از مدل واژگانی صحت بازشناسی واج را تا ۸۲/۸٪ افزایش می‌دهد.

کلمات کلیدی: بازشناسی گفتار- شبکه‌های عصبی استخراج کننده مؤلفه‌های اساسی غیرخطی- معکوس سازی شبکه‌های عصبی- شبکه‌های عصبی دوسویه- مدل سازی واژگانی

۱- مقدمه

تنوع در تلفظ[۱] که به علل تنوع در نحوه صحبت[۲]، میزان رسمیت، محیط، ناتوانی‌های گفتاری، لهجه، تفاوت‌های آناتومیکی و عوامل متعدد دیگر به وجود می‌آید، سبب ایجاد خطا در بازشناسی گفتار می‌شود. به منظور کاهش این خطا، گونه‌های مختلف از تلفظ‌های متفاوت کلمات به واژه‌نامه اضافه می‌شوند تا نرخ خارج از واژه‌نامه[۳] کاهش یابد زیرا نرخ کمتر OOV سبب کمتر شدن خطا می‌شود. روشن است که ایجاد یک واژه‌نامه تلفظ[۴]، یعنی واژه‌نامه‌ای که تلفظ‌های چندگانه یک کلمه را شامل شود، از ضرورت‌های بسیار مهم در سامانه‌های بازشناسی گفتار است. تأثیر بسزای این واژه‌نامه در افزایش کارآیی بازشناخت گفتار پیوسته دادگان بزرگ[۵] غیر قابل انکار است[۱]. واژه‌نامه‌ای که بر اساس قواعد آوایی[۶] در IBM[۲] به‌کار گرفته شد، نخستین مورد موفق در این زمینه است. اما خلق چنین واژه‌نامه‌ای به صورت دستی و یا مبتنی بر قواعد آوایی، بسیار وقت‌گیر است و از سوی دیگر نیاز به مهارت زیادی دارد[۳]. بنابراین، ضرورت ایجاد چنین واژه‌نامه‌ای به صورت خودکار تلاش‌های محققان زیادی را از اوایل دهه ۹۰ به مدل سازی تلفظ[۷] به خود جلب کرده است[۶،۵،۴]. با وجود استخراج خودکار قواعد تلفظی در این روش، نیاز به آواشناس خبره به جهت برچسب‌دهی دادگان و نیز حجم دادگان آموزشی بزرگ، حتی در صورت برچسب‌دهی خودکار از مشکلات این روش است[۱۰،۹،۸،۷]. با توجه به وجود تنوع تلفظی بیشتر در گفتار محاوره‌ای نسبت به گفتار خواندنی[۱۳،۱۲،۱۱]، در مدل‌سازی تلفظ مربوط به این دادگان روش‌های هوشمند نظیر شبکه‌های عصبی مصنوعی و نیز روش‌های آماری به‌کار گرفته می‌شود[۱۵،۱۴،۱۳]. در این روش‌ها شبکه عصبی که معمولاً شبکه تلفظ[۸] نامیده می‌شود و یا روش آماری که نقش مدل تلفظ را برعهده دارد، احتمال توالی‌های واج مختلف را به صورت خودکار استخراج می‌کند. با به‌کارگیری این احتمال در قانون تصمیم‌گیری بیز، صحت بازشناسی به صورت قابل ملاحظه‌ای افزایش نشان می‌دهد.

مقاله حاضر نیز تلاشی در جهت ایجاد مدل‌های واژگانی بر اساس شبکه‌های عصبی و پردازش‌های دوسویه است. شبکه‌های عصبی دوسویه با الهام از ایده‌های محاسباتی مغز انسان سعی در نزدیک‌تر شدن به کارآیی، انعطاف پذیری، صحت و قابلیت اطمینان سامانه درک گفتار

عصبی در مغز انسان را دارند. استفاده از این شبکه‌ها در سامانه‌های بازشناسی گفتار و در حذف تنوعات مختلف از ورودی نتایج قابل ملاحظه‌ای داشته است[۱۹،۱۸،۱۷،۱۶]. در این روش سعی می‌شود تا با مدل‌سازی دنباله واج مربوط به واژگان مختلف در شبکه عصبی از یک‌سو دنباله واج گونه‌های تلفظ‌های اشتقاقی به صورت خودکار و با استفاده از پردازش دوسویه در شبکه‌های عصبی به دنباله واج تلفظ رسمی و از سوی دیگر دنباله واج بازشناسی شده توسط مدل صوتی به دنباله واج برچسب شده تصحیح شود.

تلفظ رسمی و اشتقاقی عبارت است از:

تلفظ رسمی[۹]: دنباله واج‌های استاندارد که در گفتار خواندنی تلفظ می‌شود و تنوعاتی نظیر تنوع در گوینده، لهجه و همجواری آواها در آن در نظر گرفته نمی‌شود.

تلفظ اشتقاقی[۱۰]: دنباله واج تلفظ شده در گفتار که تلفظ‌های مختلف مربوط به تنوع گوینده و محاوره‌ای بودن گفتار را شامل می‌شود.

به این ترتیب مدل واژگانی، قادر خواهد بود تا تلفظ‌های چندگانه یک کلمه را به تلفظ رسمی آن کلمه تصحیح نماید. در این روش به دادگان آموزشی با حجم بزرگ نیاز نیست زیرا شبکه‌های عصبی قادرند تا برای ورودی آزمون که هرگز در میان دادگان آموزش نبوده است، بر طبق آنچه که از دادگان آموزش فراگرفته‌اند، تصمیم مناسبی اتخاذ نمایند[۲۰].

هم‌چنین مدل واژگانی می‌تواند تا خطاهای بازشناسی مدل صوتی را بر اساس یادگیری دنباله‌های واج مجاز، برطرف نماید.

تمامی عوامل مذکور در علت ایجاد تنوع در تلفظ، با تأثیر بر روی دادگان صوتی سبب ایجاد تلفظ‌های چندگانه در یک کلمه واحد می‌شوند. تصحیح دادگان صوتی می‌تواند در حذف تلفظ‌های چندگانه موثر باشد، زیرا علت عمده را باید در تنوع دادگان صوتی جستجو نمود. تصحیح دنباله واج با استفاده از مدل‌سازی واژگانی مبتنی بر پردازش‌های دوسویه شبکه‌های عصبی و معکوس‌سازی این شبکه‌ها، این امکان را فراهم می‌سازد تا با دو سویه کردن پردازش از سطح واژگان تا سطح دادگان صوتی(پارامترهای بازنمایی) و اصلاح این پارامترها، صحت بازشناسی را مجدداً افزایش داد.

در بخش دوم مقاله حاضر به تحلیل شبکه‌های عصبی مؤلفه‌های اساسی غیرخطی و در بخش سوم به روش‌های معکوس‌سازی شبکه‌های عصبی پرداخته‌می‌شود. در بخش چهارم دادگان مورد استفاده و نحوه استخراج ویژگی از دادگان بیان می‌شود. در بخش پنجم مدل صوتی استفاده شده تشریح می‌شود. در بخش ششم مدل‌های واژگانی طراحی شده بیان می‌شود و در بخش هفتم به بیان دو روش به منظور تلفیق دانش واژگانی با اطلاعات صوتی با استفاده از روش‌های معکوس‌سازی شبکه‌های عصبی و در بخش بعدی به جمع بندی نتایج حاصل می‌پردازیم.

۲- تحلیل مؤلفه‌های اساسی خطی و غیرخطی توسط شبکه‌های عصبی[۲۱]

تحلیل مؤلفه‌های اساسی روشی برای نگاشت دادگان چندبعدی به فضایی با بعد کمتر و با کمترین اتلاف اطلاعات است. تحلیل مؤلفه‌های اساسی غیرخطی[۱۱] از روش‌های جدید برای تحلیل دادگان چند بعدی، مشابه روش شناخته‌شده تحلیل مؤلفه‌های اساسی[۱۲] است. از کارآیی‌های تحلیل مؤلفه‌های اساسی خطی و غیرخطی حذف نویز و تنوعات از سیگنال ورودی است به طوری‌که اگر تنوعات مذکور در راستای مؤلفه‌های استخراج شده تصویر نداشته باشند در بازسازی، تنوعات و نویز همراه ورودی تا حد قابل قبولی حذف خواهد شد. در تحلیل مؤلفه‌های اساسی برای اجتناب از محاسبات پیچیده و حجیم، روش‌های هوشمند نظیر شبکه‌های عصبی جایگزین روش‌های آماری ، ترکیبی و یا ساختاری می‌شوند.

در $NLPCA$ ، نگاشت به فضای ویژگی و تبدیل معکوس آن یک نگاشت غیرخطی مطابق روابط زیر است.

$$\underline{T}=\underline{G}(\underline{Y}) \qquad (۱)$$

$$\underline{Y'}=\underline{H}(\underline{T}) \qquad (۲)$$

توابع $\underline{G}$ و $\underline{H}$ در جهت کمینه‌کردن خطا به‌دست می‌آیند. در به‌دست آوردن $\underline{G}$و $\underline{H}$ و به منظور اجتناب از محاسبات پیچیده، از شبکه‌های عصبی مطابق شکل۱ استفاده می‌شود. شبکه مذکور از طریق آموزش با سرپرستی قابل آموزش است.

شبکه عصبی NLPCA محدود به ساختار فوق نیست و با افزایش تعداد لایه‌های شبکه به هفت لایه، امکان تقریب ناپیوستگی‌های احتمالی در توابع $\underline{G}$ و$\underline{H}$ فراهم و لذا قابلیت‌های بالاتری نسبت به شبکه پنج لایه کرامر[۱۳]ممکن می‌شود.

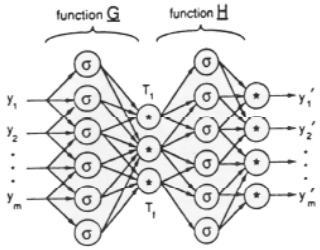
۳- روش‌های معکوس سازی شبکه‌های عصبی در تشکیل شبکه‌های عصبی دوسویه[۱۴]

در این قسمت دو روش برای معکوس سازی شبکه‌های عصبی و به دست آوردن ورودی از خروجی معرفی می‌کنیم. در روش اول پس از آموزش شبکه با دادگان آموزش، این امکان وجود دارد که با اعمال یک ورودی دلخواه به شبکه و به دست آوردن خروجی آن، بردار خطای ما بین خروجی مطلوب و خروجی فعلی محاسبه شود و بدون اصلاح وزن‌ها، خطا به لایه‌های قبل پس انتشار شود تا در لایه ورودی، ورودی اصلاح شود[۲۲]. در این روش وزن‌ها ثابت و ورودی همانند وزن‌ها اصلاح می‌شود تا ورودی بهینه از دیدگاه شبکه حاصل شود. روش دوم بر مبنای آموزش دو شبکه معکوس یکدیگر بر روی دادگان آموزش است، به این معنا که برای شبکه جلوسو(مستقیم) ورودی و خروجی به ترتیب ورودی و خروجی دادگان آموزش و در شبکه معکوس، عکس آن است[۲۳]. در ادامه با بررسی‌های تحلیلی نشان خواهیم داد که هر دو روش قادر به به‌دست آوردن ورودی از خروجی هستند.

۳-۱ معکوس سازی شبکه با اصلاح ورودی به روش گرادیان

در این روش، از پس انتشار خطا به منظور اصلاح ورودی استفاده می‌شود. یک شبکه جلوسو با یک لایه پنهان H و لایه‌های ورودی و خروجی IوO مفروض است. x_k^t مؤلفه k ام بردار ورودی در تکرار t که از x_k^0 آغاز شده است، در روش گرادیان، مطابق رابطه زیر محاسبه می‌شود.

$$x_k^{t+1} = x_k^t - \eta \frac{\partial E}{x_k^t} \qquad k \in I, t = 0,1,2,... \qquad (۳)$$



شکل ۱- معماری شبکه برای به‌دست آوردن fفاکتور غیر خطی با بکار گیری شبکه خود انجمنی(σ گره‌های سیگموییدی و * گره‌های سیگموییدی یا خطی را نشان می دهد.)

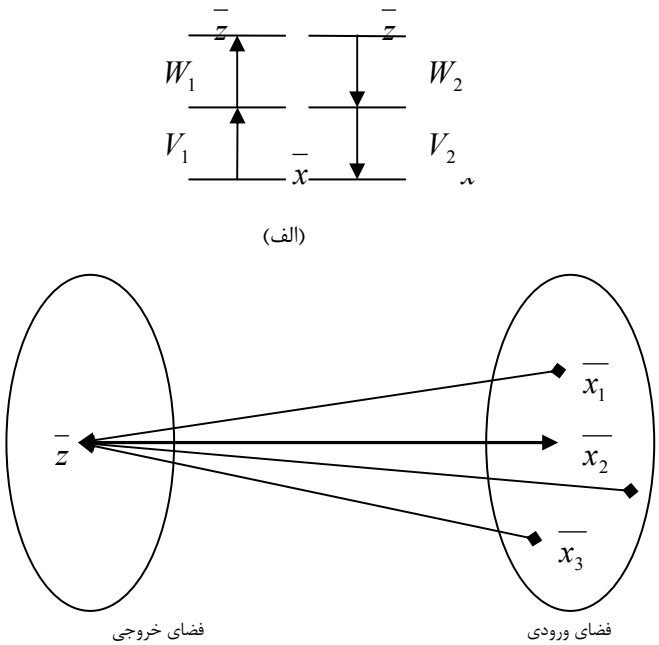
در رابطه فوق، گرادیان خطا بر اساس قانون دلتا به‌صورت زیر به‌دست می‌آید.

$$\delta_j = \begin{cases} f_j'(y_j)(y_j - d_j) & j \in O \\ f_j'(y_j) \sum_{m \in H,O} \delta_m w_{jm} & j \in I,H \end{cases} \quad (۴)$$

در این رابطه f_j تابع غیرخطی نورون j ام، y_j میزان فعالیت نورون j ام، d_j خروجی مطلوب و w_{jm} وزن اتصال بین نورون j و نورون m ام است. این روش مانند سایر روش‌های مبتنی بر گرادیان، احتمال گیرافتادن در کمینه‌های موضعی را دارد.

۲-۳ معکوس سازی شبکه به روش آموزش شبکه معکوس

روش دیگری که در شبکه‌های عصبی برای به‌دست آوردن ورودی از خروجی وجود دارد، استفاده از دو شبکه جلوسو به صورت معکوس یکدیگر است. آموزش این شبکه‌ها در صورتی که دادگان آموزش یک به یک باشند هم‌گرا خواهد شد. به عنوان نمونه اگر دادگان چند به یک باشند شبکه مستقیم هم‌گرا می‌شود و شبکه معکوس میانگین ورودی‌ها را بازسازی می‌کند[۲۴]. این شبکه‌ها می‌توانند بر اساس الگوریتم پس انتشار خطا جداگانه و یا در قالب یک الگوریتم دوسویه آموزش داده شود[۲۴]. ساختار کلی دو شبکه معکوس، هر کدام به عنوان نمونه با تک لایه پنهان در شکل ۲- (الف) نشان داده شده است. شکل ۲- (ب) نگاشت بین فضای ورودی و خروجی در این شبکه‌ها را نمایش می‌دهد.



شکل ۲- (الف) دو شبکه تک لایه پنهان معکوس یکدیگر، (ب) نگاشت بین فضای ورودی و خروجی

۴- دادگان گفتاری و استخراج بازنمایی

۴-۱ دادگان گفتاری

به علت نیاز به وجود تنوعات بالا در دادگان این تحقیق، دادگان گفتاری که مورد استفاده قرار می‌گیرد، دادگان فارس دات تلفنی[۱۵] است. این دادگان از ۵ بخش تشکیل شده است که توسط ۶۴ گویشور بیان شده است. از میان تمامی کلمات بیان شده در دادگان، ۴۰۰ کلمه که حداقل بیش از یک‌بار ادا شده‌اند از ۶۴ گویشور انتخاب و در آموزش و آزمون شبکه‌ها مورد استفاده قرار می‌گیرد[۲۵]. ۷۵٪ از دادگان انتخابی، به عنوان دادگان آموزش و ۲۵٪ از آن به عنوان دادگان آزمون، در آموزش و آزمون تمامی شبکه‌ها استفاده می‌شود. بنابراین در کلماتی که توسط تمامی گویشوران ادا شده است، دادگان مربوط به ۴۸ گویشور به جهت آموزش و ۱۶ گویشور در آزمون مدل‌ها استفاده شده است. نرخ نمونه برداری دادگان ۱۱۰۲۵ هرتز است که به ۸ کیلو هرتز(نرخ نمونه برداری گفتار تلفنی) تغییر داده می‌شود. پنجره‌گذاری به جهت استخراج ویژگی از دادگان، با طول پنجره‌ی ۱۲۸ و با گام پیشروی ۶۴ نمونه، انجام می‌شود.

۴-۲ استخراج پارامترهای بازنمایی

در اکثر روش‌های استخراج ویژگی طیفی، مبتنی بر تبدیل فوریه، پس از محاسبه طیف توسط تبدیل فوریه گسسته زمان کوتاه، از یک بانک فیلتر جهت محاسبه میزان انرژی طیف در باندهای فرکانسی متفاوت استفاده می‌گردد. در این میان روش‌های الهام گیرنده از سامانه‌های زیستی عملکرد بهتری از خود نشان می‌دهند[۲۶و۲۷]. مهم‌ترین روشی که از سامانه شنوایی انسان در استخراج ویژگی اقتباس شده است، انتخاب مقیاس غیرخطی برای محور فرکانس است. در میان روش‌های متداولی که بر اساس دو نوع مقیاس غیرخطی بارک[۱۶] و مل[۱۷] ارایه شده است ضرایب کپستروم مقیاس مل(MFCC[۱۸]) از متداول‌ترین روش‌های موجود در استخراج پارامترهای بازنمایی است که در آن پس از حصول مقادیر لگاریتم انرژی بانک فیلترها، این مقادیر به حوزه کپستروم انتقال می‌یابند. به منظور هنجارسازی[۱۹] پارامترهای بازنمایی، هر پارامتر به تغییراتی که در کل دادگان مربوط به هر گوینده دارد بهنجار می‌شود(هنجارسازی طولی پارامترهای بازنمایی، جداگانه برای هر گوینده). مزیت این روش نسبت به هنجارسازی طولی در کل دادگان این است که پارامترها تا حدی نسبت به گوینده هنجارسازی می‌شوند. به این روش، هنجارسازی به میانگین و واریانس نیز گفته می‌شود. در ادامه روابط این روش هنجارسازی آورده شده است.

$$\bar{k}_1 = \frac{1}{M} \sum_{i=1}^{M} \bar{x}_i \quad (۵)$$

$$\bar{q}_i = \bar{x}_i - \bar{k}_1 \quad (۶)$$

$$\bar{k}_2 = \frac{1}{M} \sum_{i=1}^{M} \bar{q}_i^{\,2} \quad (۷)$$

$$\bar{x}_i^{\,normalized} = \left\{ \frac{q_{in}}{\sqrt{k_{2n}}}, n = 1,2,...,k,...,N \right\} \quad (۸)$$

در روابط فوق M تعداد بردارهای بازنمایی در کل دادگان مربوط به یک گوینده و N تعداد پارامترهای هر بردار بازنمایی است. با اضافه کردن مشتقات مرتبه اول و دوم(ضرایب مرتبه اول و دوم دلتا)، به منظور هنجارسازی بهتر نسبت به گوینده و نیز به جهت افزودن اطلاعات در رابطه با ویژگی‌های پویای سیگنال، تعداد پارامترهای هر بردار بازنمایی(N) به ۳۹ پارامتر می‌رسد(با توجه به محدوده ۱۲۵ تا ۳۷۰۰ هرتزی برای دادگان تلفنی از بین ۱۸ بانک فیلتر طراحی شده برای گفتار غیرتلفنی، فیلترهای دوم تا چهاردهم انتخاب و با محاسبه لگاریتم انرژی این ۱۳ فیلتر، بردار بازنمایی اصلی به‌دست می‌آید). ضریب مرتبه i ام دلتا به صورت زیر به‌دست می‌آید.

$$\Delta^{i}\left\{\bar{u}_t\right\} = \Delta^{i-1}\left\{\bar{u}_{t+1}\right\} - \Delta^{i-1}\left\{\bar{u}_{t-1}\right\} \quad , \quad \Delta^{0}\left\{\bar{u}_t\right\} = \bar{u}_t \qquad (۹)$$

۵- مدل صوتی[۲۰]

مدل صوتی که استخراج‌کننده آوای متناظر هر فریم، براساس پارامترهای بازنمایی است؛ یک شبکه MLP با دو لایه پنهان، با ساختار نورونی ۳۹×۹-۲۵۶-۱۲۸-۳۴ است که در ورودی ۹ فریم متوالی(۴ فریم مجاور راست و ۴ فریم مجاور چپ) قرار داده می‌شوند، تا در خروجی فریم میانی توسط ۳۴ نورون (به تعداد آواها) بیان شود. توابع غیر خطی لایه‌های پنهان و نیز خروجی به جهت تسریع بیشتر در هم‌گرایی شبکه، تانژانت هیپربولیک[۲۱] با مقادیر خروجی بین ۱و ۱- انتخاب می‌شود. وزن‌های اولیه بر طبق الگوریتم نگوین-ویدرو[۲۲] انتخاب می‌شود. شبکه با ضریب یادگیری و ضریب ممنتوم متغیر تا هم‌گرایی کامل و بر اساس الگوریتمRPROP[۲۳] آموزش می‌بینند. این الگوریتم علاوه بر سرعت مطلوب در هم‌گرایی، خطای مناسبی نیز دارد[۲۸]. پس از آموزش مدل، در مرحله آزمون، صحت بازشناسی فریم، بر روی فریم‌های غیرسکوت ۷۱/۳۴٪ به‌دست می‌آید.

در تبدیل فریم به واج آن‌چه مد نظر قرار می‌گیرد، متوسط طول واج‌ها بر حسب فریم است. چون به‌طورمعمول طول واج‌ها در دادگان از دو فریم بیشتر است، تبدیل به صورت زیر انجام خواهدشد. در دنباله فریم به‌دست آمده از خروجی مدل‌صوتی، اگر تعداد فریم‌های متوالی از یک نوع، بیشتر یا برابر ۲ باشد به عنوان واج در نظر گرفته می‌شود. بنابراین از فریم‌های تکی صرف‌نظر می‌گردد. میزان صحت[۲۴] و دقت[۲۵] واج بازشناسی شده مدل صوتی از دادگان آزمون به همراه میزان حذف[۲۶]، درج[۲۷] ، جانشینی[۲۸] در جدول ۱ مشاهده می‌شود[۲۹].

جدول۱- نتایج بازشناسی واج دادگان آزمون مدل صوتی مرجع

میزان دقت بازشناسی واج	میزان صحت بازشناسی واج	میزان حذف واج	میزان درج واج	میزان جانشینی واج
٪۶۱/۱	٪۷۱/۶	٪۱۳/۷	٪۱۰/۴	٪۱۴/۷

۶- مدل‌های واژگانی

تا کنون شبکه‌های عصبی جلوسوی چندلایه، به طوروسیعی جهت بازشناسی الگوها مورد استفاده قرار گرفته‌اند. با این همه ساختار یک سویه شبکه‌های عصبی جلوسو، در بازشناسی الگوها به‌خصوص مواردی که الگوها تحت تأثیر نویز قرار گرفته‌اند و یا در اثر عوامل مختلف تغییر یافته‌اند که به آن تنوعات[۳۰] در الگوهای ورودی و یا عدم انطباق دادگان[۳۱] آموزش با آزمون گفته می‌شود، نارسا هستند[۲۴]. از سوی دیگر شواهد موجود حاکی از پردازش دوسویه الگوهای مورد بازشناسی در مغز هستند[۲۹]. برای افزایش کارآیی، رویکرد دو سویه‌سازی پردازش در شبکه‌های عصبی، رویکرد مناسبی به نظر می‌رسد. از سوی دیگر،شیوه پردازش مقاوم الگوها در مغز انسان خصوصاً در شرایطی که در الگوهای ورودی، تنوعات وجود داشته باشد، توانایی مناسب جداسازی از طریق مؤلفه‌های اساسی غیرخطی سیگنال‌ها در مغز را نشان می‌دهد[۲۴].

یافته‌های تحقیقات علوم اعصاب نیز حاکی از وجود اتصالات دو سویه و تعامل دو طرفه مابین نواحی اولیه و نواحی بالاتر پردازش‌های حسی در قشر مغز هستند. به نظر می‌رسد که قشر مخ در پاسخ به ورودی‌های حسی، علاوه بر پردازش‌های از پایین به بالا از طریق اعصاب آوران، از طریق پردازش‌های بالا به پایین توسط اعصاب وابران به سیگنال‌های حسی ورودی، دانش اضافه می‌کند تا فرضیه‌ها و تصورهای بازسازی شده به واقعیت اشیاء در جهان واقعی تا حد ممکن نزدیک‌تر گردند[۳۱،۳۰،۲۹].

شکل‌گیری جاذب‌ها[۳۲]در قعرهای بستر جذب‌ها[۳۳] از خصوصیات شبکه‌های عصبی با اتصالات دوسویه است که ترکیب این خواص با توانایی‌های شبکه‌های جلوسو، امکان طراحی مدل‌های مفیدتر را فراهم می‌سازد[۲۴].

در ادامه با طرح مدل‌های دوسویه شبکه‌های عصبی، که با استفاده از معکوس‌سازی شبکه‌های عصبی قابل پیاده‌سازی هستند، توانایی‌های شبکه‌های عصبی به جهت رفع تنوعات واژگانی در یک سامانه بازشناسی گفتار نشان داده می‌شود. دو هدف کلی در تمامی مدل‌های طراحی شده مدنظر است:

الف- اصلاح دنباله واج خروجی مدل صوتی در قیاس با دنباله واج برچسب شده.

ب- تصحیح دنباله واج تلفظ‌های چندگانه در خروجی مدل صوتی به دنباله واج تلفظ رسمی.

مدلNLPCA : در این روش یک شبکه عصبیMLP خودنگاشت با سه لایه پنهان با توابع غیرخطی تانژانت هیپربولیک و یک لایه خروجی با تابع خطی به‌کار گرفته می‌شود. این شبکه همان‌گونه که در بخش دوم بیان شد، یک شبکه NLPCA است که یک‌بار جهت نگاشت دنباله واج در تلفظ‌های چندگانه یک کلمه به دنباله واج تلفظ رسمی آن کلمه و بار دیگر جهت نگاشت دنباله واج در تلفظ‌های چندگانه یک کلمه به خود همان دنباله به‌کار گرفته می‌شود. علت استفاده از دو دسته دادگان آموزش مختلف در این شبکه، به ترتیب اصلاح دنباله واج در قیاس با دنباله واج تلفظ رسمی و برچسب شده است. در جدول ۲ به‌عنوان نمونه، برخی از تلفظ‌های اشتقاقی و تلفظ رسمی کلمه "چهارده" نشان داده شده‌است.

جدول۲- تلفظ رسمی و برخی از تلفظ‌های اشتقاقی رایج کلمه "چهارده"[۳۴]

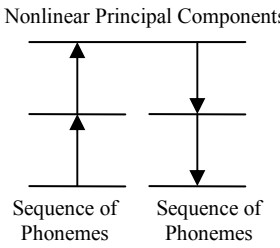
تلفظ رسمی	برخی از تلفظ‌های اشتقاقی
$'/h/r=dah	$'/r=dah
	$'/r=da
	$'ah/r=dah
	$'/h/rdah
	$'/h/r=deh
	$'/h/r=da

در ورودی و خروجی این مدل هر واج با ۳۴ بیت بیان می‌شود. این نحوه بیان، قابلیت اتصال مستقیم مدل مذکور به لایه‌های پایین‌تر را فراهم می‌سازد(مدل صوتی در این‌جا). با توجه به مشخص‌بودن مرز کلمات در این مدل(کلمات مجزا[۳۵])، ورودی این شبکه متناظر با دنباله واج طویل‌ترین کلمه در نظر گرفته می‌شود. برای کلمات کوچک‌تر از این طول، در انتهای کلمه، واج سکوت قرارداده می‌شود. با توجه به دادگان مورد استفاده در آموزش و آزمون این مدل، ابعاد شبکه۳۴×۱۶-۱۲۸-۶۴-۱۲۸-۱۶×۳۴ در نظر گرفته می‌شود(ابعاد ورودی و خروجی معرف آن است که در دادگان استفاده شده، طویل‌ترین کلمه از ۱۶ واج تشکیل شده است). شکل۳ ساختار این شبکه را نشان می‌دهد. با دقت در این شبکه و با توجه به نحوه ترسیم روشن است که این شبکه از یک شبکه جلوسوی مستقیم و یک شبکه معکوس تشکیل یافته است که نورون‌های لایه پنهان میانی از یک‌سو خروجی شبکه مستقیم و ورودی شبکه معکوس و از سوی دیگر مؤلفه‌های اساسی غیرخطی کلمات مختلف هستند.

پس از مقداردهی اولیه وزن‌ها، شبکه مذکور مطابق روش مطرح شده در مورد مدل صوتی، آموزش داده می‌شود. در آموزش شبکه از کلمات ادا شده توسط ۴۸ گویشور استفاده می‌شود. در آزمون مدل، دنباله واج استخراج شده از دنباله فریم بازشناسی شده در خروجی مدل صوتی مرجع از دادگان آزمون، به ورودی مدل اعمال می‌شود. با توجه به نحوه استخراج واج از فریم، میانگینی از فریم‌های موجود در یک واج محاسبه و دنباله واج استخراج می‌شود. نتایج حاصل، در جدول۳ مشاهده می‌شود. لازم به ذکر است که نتایج جداول ۳- (الف) و ۳- (ب)، به ترتیب با در نظر گرفتن دنباله واج تلفظ رسمی کلمه و دنباله واج برچسب شده در

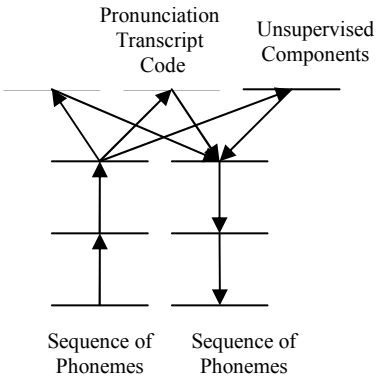
دادگان، به عنوان خروجی مطلوب محاسبه شده‌است. نتایج به‌دست آمده، ارتقاء نتایج مدل صوتی مرجع درقیاس با تلفظ رسمی و نیز تلفظ‌های برچسب شده را نشان می‌دهد.

از آن‌جا که شبکه NLPCA مشابه یک فیلتر غیرخطی عمل می‌کند[۲۴] و می‌تواند دنباله واج ورودی را با توجه به اطلاعات واژگانی که درون آن شکل گرفته است فیلتر و تصحیح نماید؛ و به منظور افزایش کارآیی مدل با نظر به اینکه شبکه معکوس، ورودی شبکه مستقیم را در خروجی خود تولید می‌نماید، اعمال خروجی به‌دست آمده از شبکه معکوس به ورودی شبکه مستقیم و به‌دست آوردن مجدد خروجی شبکه مستقیم و تکرار این عمل تا عدم تغییر در خروجی شبکه معکوس، سبب افزایش قابل توجهی در کارآیی این مدل می‌شود که نتایج آورده شده در سطرهای دوم و سوم از جداول۳- (الف) و ۳- (ب) مؤید این مطلب است.



شکل۳- مدل NLPCA

مدلNLPCA با سرپرستی : همان‌گونه که در تشریح مدل قبلی بیان شد، در لایه پنهان میانی مدلNLPCA، مؤلفه‌های اساسی غیرخطی متناظر ورودی استخراج می‌شود. دنباله‌های واج مختلف نظیر یک کلمه، از جهت بیان نمودن یک کلمه واحد، اشتراک و به لحاظ تفاوت در تلفظ آن کلمه، اختلاف دارند. لذا با جداسازی وجوه اشتراک و تمایز در نورون‌های لایه پنهان میانی، انتظار کارآیی بالاتری از مدل می‌رود. به این منظور تعدادی از نورون‌های لایه میانی مدل قبلی (۹ نورون) را به کد باینری کلمه[۳۶] بیان شده در ورودی (حداکثر ۵۱۲ کلمه) و بخش دیگری را به کد نسخه تلفظی[۳۷] نظیر دنباله ورودی اختصاص می‌دهیم(شکل ۴). تعداد نورون‌های دسته دوم، بسته به حداکثر تعداد نسخ تلفظی مربوط به کلیه کلمات دادگان است. در دادگان مورد استفاده، با وجود حداکثر ۳۲ گونه تلفظی متفاوت برای هر کلمه، ۵ نورون به بیان باینری کد نسخه تلفظی اختصاص می‌یابد.



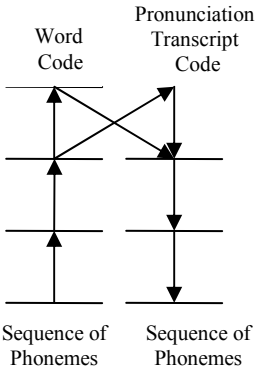
شکل۴- شبکه مربوط به مدل NLPCA با سرپرستی

توابع غیرخطی تمامی لایه‌های پنهان، تانژانت هیپربولیک و لایه خروجی خطی در نظر گرفته می‌شود. این شبکه تا رسیدن به خطای مطلوب مطابق آن‌چه که قبلاً در مورد مدل NLPCA و نیز مدل صوتی مرجع تشریح شد، جهت نگاشت دنباله واج نظیر هر کلمه به خود همان دنباله، آموزش داده می‌شود. آزمون این مدل به‌مانند مدل قبلی، با خروجی به‌دست آمده از مدل صوتی انجام می‌گیرد. در آزمون مدل کنونی، خروجی به‌دست آمده از مدل صوتی بر روی دادگان آزمون، به ورودی شبکه اعمال می‌شود. پس از به‌دست آمدن کد نسخه تلفظی، کد نسخه تلفظی به‌دست آمده به کد نظیر نسخه رسمی آن کلمه تغییر می‌یابد، این کد که در آموزش شبکه‌ها نیز متناظر با نسخه رسمی در نظر گرفته شده‌است، به عنوان نمونه می‌تواند ۱۱۱۱۱ باشد. پس از تغییر کد نسخه تلفظی، خروجی شبکه به‌دست می‌آید. نتایج جدول ۳ با در نظر گرفتن دو خروجی متفاوت به عنوان خروجی مطلوب به‌دست آمده‌است. نتیجه جدول ۳- (الف) بدون تغییر کد نسخه تلفظی و جدول ۳- (ب) با تغییر این کد به کد نظیر تلفظ رسمی محاسبه شده است.

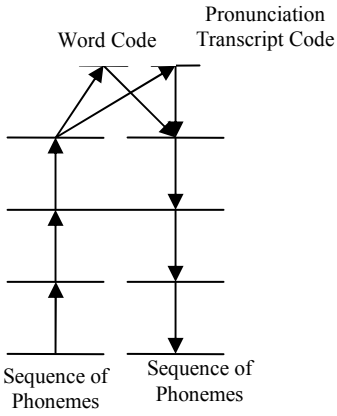
مدل شبکه معکوس: همان‌طور که در مورد دو مدل قبلی نیز بیان‌شد، شبکه NLPCA از نگاهی‌دیگر، از ترکیب دو شبکه جلوسوی مستقیم و معکوس، تشکیل شده است. در بخش سوم و در بیان روش‌های معکوس‌سازی شبکه‌های عصبی، چگونگی حصول ورودی بهینه توسط این شبکه‌ها تشریح شد. به منظور استفاده از توانایی‌های شبکه معکوس در مدل‌سازی واژگانی، در مدل قبلی با حذف نورون‌های مربوط به مؤلفه‌های بدون سرپرستی از لایه میانی مطابق شکل ۵ ساختار مدل به دو شبکه معکوس یکدیگر تغییر می‌یابد. در خروجی شبکه مستقیم و ورودی شبکه معکوس، کد باینری کلمه و نسخه تلفظی آن قرار می‌گیرد. در آموزش این زوج شبکه، دنباله واج تلفظ‌های مختلف کلمات ادا شده توسط ۴۸ گویشور، در ورودی شبکه مستقیم و خروجی شبکه معکوس قرار می‌گیرد. خروجی شبکه مستقیم و ورودی شبکه معکوس که به کد باینری کلمه و نسخه تلفظی آن اختصاص دارد، به‌مانند مدل قبلی شامل ۱۴ نورون است. آموزش این دو شبکه می‌تواند به صورت کاملاً مجزا و به مانند روش‌های قبلی انجام شود.

در آزمون این مدل به‌مانند مدل قبلی، خروجی به‌دست آمده از مدل صوتی بر روی دادگان آزمون، به ورودی شبکه مستقیم اعمال می‌شود. پس از به‌دست آوردن خروجی شبکه مستقیم،کد نسخه تلفظی به‌دست آمده به کد نظیر نسخه رسمی آن کلمه تغییر می‌یابد. این کد که در آموزش شبکه‌ها نیز متناظر با نسخه رسمی در نظر گرفته شده‌است، به عنوان نمونه می‌تواند ۱۱۱۱۱ باشد. با تغییر کد نسخه تلفظی در خروجی شبکه مستقیم، کد کلمه به همراه کد نسخه تلفظی تغییر یافته در ورودی شبکه معکوس قرار می‌گیرد تا خروجی این شبکه به‌دست آید. نتایج آزمون این مدل بدون تغییر کد نسخه تلفظی در جدول ۳- (الف)، و با تغییر این کد، در جدول ۳- (ب) مشاهده می‌شود. نتایج جدول ۳- (ب) حاکی از آن است که شبکه معکوس تا حد مطلوبی، قادر به تصحیح دنباله واج تلفظ‌های اشتقاقی به تلفظ رسمی هر کلمه است.

مدلNLPCA با سرپرستی توسط شبکه معکوس: شکل‌گرفتن بدون سرپرستی دانش واژگانی در لایه مؤلفه‌های اساسی غیرخطی در مدل NLPCA، معرف توانایی‌های بالای این شبکه در مدل‌سازی واژگانی است. از سوی دیگر قابلیت‌های شبکه معکوس در بازسازی ورودی از خروجی را نیز نمی‌توان نادیده گرفت. به منظور تلفیق مزایای این دو مدل، ساختار شکل ۶ استفاده می‌شود. این مدل در حقیقت یک شبکه NLPCA است که در آن شبکه جلوسوی مستقیم و معکوس بر روی لایه مؤلفه‌های اساسی غیرخطی اعمال می‌شود. ساختار نورونی این مدل مشابه مدلNLPCA است و ابعاد شبکه جلوسوی مستقیم و معکوس ۶۴-۳۲-۱۴ است. تمامی توابع غیرخطی تانژانت هیپربولیک و تنها در لایه خروجی شبکه NLPCA خطی است. نتایج آزمون این مدل با استفاده از دادگان آزمون مدل‌های قبلی بدون تغییر وبا تغییر کد نسخه تلفظی به ترتیب در جداول ۳- (الف) و ۳-(ب) نشان داده شده است. شکل ۷ نمودار نتایج جدول ۳ است.



شکل۵- شبکه مربوط به مدل شبکه معکوس



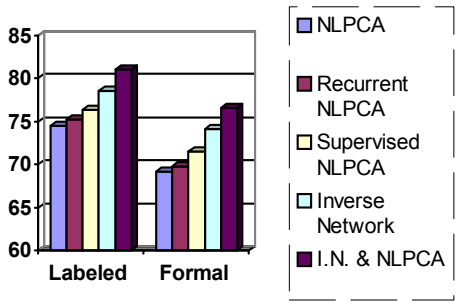
شکل۶- شبکه مربوط به مدل NLPCA با سرپرستی توسط شبکه معکوس

جدول۳- (الف) نتایج اصلاح واج دادگان آزمون مدل‌های واژگانی با در نظر گرفتن دنباله واج برچسب شده به عنوان خروجی مطلوب

مدل	میزان دقت بازشناسی واج	میزان صحت بازشناسی واج	میزان حذف واج	میزان درج واج	میزان جانشینی واج
مدل NLPCA	٪۶۷	٪۷۴/۵	٪۱۲	٪۷/۵	٪۱۳/۵
مدلNLPCA با تکرار مجدد خروجی در ورودی	٪۶۷/۹	٪۷۵/۲	٪۱۱/۷	٪۷/۳	٪۱۳/۲
مدل NLPCA با سرپرستی بدون تغییر کد نسخه تلفظی	٪۶۹/۵	٪۷۶/۳	٪۱۱/۱	٪۶/۸	٪۱۲/۷
مدل شبکه معکوس بدون تغییر کد نسخه تلفظی	٪۷۲/۳	٪۷۸/۶	٪۹/۶	٪۶/۲	٪۱۱/۷
مدل NLPCA با سرپرستی توسط شبکه معکوس بدون تغییر کد نسخه تلفظی	٪۷۶/۵	٪۸۱	٪۸/۶	٪۴/۵	٪۱۰/۴

جدول۳- (ب) نتایج اصلاح واج دادگان آزمون مدل‌های واژگانی با در نظر گرفتن دنباله واج تلفظ رسمی به عنوان خروجی مطلوب

مدل	میزان دقت بازشناسی واج	میزان صحت بازشناسی واج	میزان حذف واج	میزان درج واج	میزان جانشینی واج
مدل صوتی مرجع و بدون استفاده از مدل واژگانی	٪۵۰/۲	٪۶۵	٪۱۶/۶	٪۱۴/۸	٪۱۸/۴
مدل NLPCA	٪۵۷/۸	٪۶۹/۲	٪۱۴/۶	٪۱۱/۴	٪۱۶/۲
مدلNLPCA با تکرار مجدد خروجی در ورودی	٪۵۸/۹	٪۶۹/۸	٪۱۴/۱	٪۱۱	٪۱۶/۱
مدل NLPCA با سرپرستی	٪۶۰/۷	٪۷۱/۵	٪۱۳/۴	٪۱۰/۸	٪۱۵/۱
مدل شبکه معکوس با تغییر کد نسخه تلفظی	٪۶۴/۳	٪۷۴/۱	٪۱۱/۸	٪۹/۸	٪۱۴/۱
مدل NLPCA با سرپرستی توسط شبکه معکوس با تغییر کد نسخه تلفظی	٪۶۹	٪۷۶/۶	٪۱۰/۶	٪۷/۶	٪۱۲/۸



شکل۷- نمودار صحت بازشناسی واج حاصل از نتایج آزمون مدل‌های واژگانی

۷- تلفیق دانش واژگانی با اطلاعات صوتی

۷-۱ ضرورت تلفیق دانش واژگانی با اطلاعات صوتی

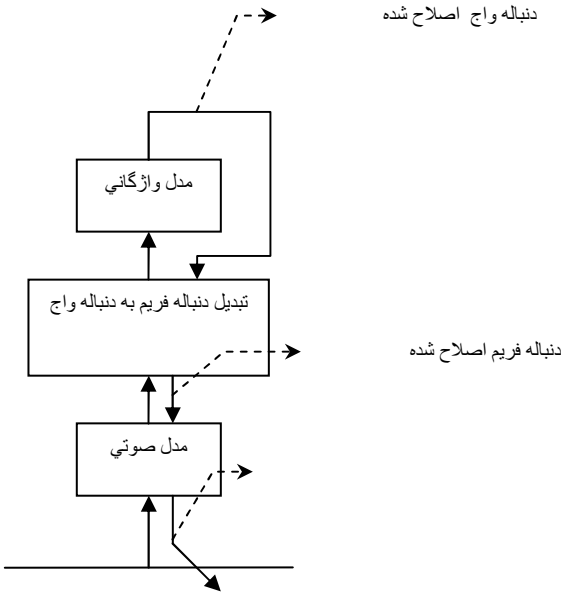
نتایج به‌دست آمده از مدل‌های واژگانی مطرح شده از دو جهت مختلف می‌تواند سبب ارتقای یک سامانه بازشناسی گفتار شود. از یک‌سو با افزایش صحت بازشناسی واج در مقایسه با تلفظ رسمی، حجم واژه‌نامه مورد نیاز برای در برداشتن تلفظ‌های چندگانه کلمات کاهش می‌یابد. لذا با واژه‌نامه کوچک‌تر، صحت بازشناسی کلمه هم‌چنان در حد قابل قبولی باقی می‌ماند. از سوی دیگر می‌تواند با تصحیح دنباله واج سبب ارتقاء صحت بازشناسی واج مدل صوتی مرجع شود. زیرا مقایسه دنباله واج تولیدشده در خروجی مدل واژگانی با دنباله واج برچسب شده در دادگان نیز حکایت از یک افزایش۹/۴۱٪ دارد. با توجه به موضوع اخیر اکنون این سؤال مطرح می‌شود که با وجود آموزش کافی مدل صوتی با استفاده از واج برچسب شده در دادگان، علت افزایش صحت واج اخیر چیست؟ بخشی از پاسخ این سؤال را بایستی در تنوعات دادگان صوتی جستجو نمود. این تنوعات که به علل مختلف در سطح دادگان صوتی ظاهر می‌شوند، علی‌رغم آموزش کافی مدل

صوتی، می‌تواند سبب خطا در بازشناسی واج مدل صوتی شود. حال اگر اختلاف ایجادشده میان دنباله واج خروجی مدل واژگانی و ورودی آن(خروجی مدل صوتی)، تنها در دادگان صوتی جستجو شود، تلفیق اطلاعات نهفته در این اختلاف با دادگان صوتی در ورودی مدل صوتی می‌تواند مفید باشد. به این منظور به کمک روش‌های مطرح شده در بخش سوم، اطلاعات سطوح بالاتر با اطلاعات سطوح پایین تلفیق می‌شود. در ادامه با به‌کارگیری دو روش بخش سوم، یعنی معکوس‌سازی به روش گرادیان و شبکه معکوس که در این بخش به ترتیب معکوس سازی با اصلاح ورودی به روش گرادیان و معکوس سازی شبکه به روش آموزش شبکه معکوس نامیده می‌شود، نتایج قابل ملاحظه‌ای در افزایش صحت بازشناسی واج حاصل می‌شود.

۷-۲ تلفیق با استفاده از معکوس سازی با اصلاح ورودی به روش گرادیان

در این روش پس از استخراج دنباله واج از دنباله فریم خروجی مدل صوتی مرجع، با استفاده از روش مطرح شده در بخش ششم، از مدل واژگانی NLPCA با سرپرستی توسط شبکه معکوس به جهت تصحیح دنباله واج به‌دست آمده به دنباله واج برچسب شده استفاده می‌شود. با مقایسه دنباله واج خروجی مدل واژگانی و دنباله واج ورودی آن (خروجی مدل صوتی) و پس انتشار خطای این دو به سطح دادگان صوتی(این روش معکوس سازی با اصلاح ورودی به روش گرادیان مدل صوتی است. به این ترتیب که دنباله واج به‌دست آمده از مدل واژگانی به عنوان خروجی مطلوب در نظر گرفته شده با پس انتشار خطای مابین خروجی مطلوب و خروجی فعلی به سطح ورودی، ورودی به روش گرادیان خطا اصلاح می‌شود) دادگان صوتی با استفاده از خطای پس انتشار شده تغییر می‌یابند.

به این ترتیب تمامی دادگان آزمون(پارامترهای بازنمایی) اصلاح می‌شوند. حال اگر مدل صوتی جدیدی که بر روی دادگان تعلیم اصلاح شده، آموزش دیده است به همراه مدل واژگانی مذکور بر روی دادگان صوتی آزمون اصلاح شده، به‌کار گرفته شود، ارتقاء نتایج بازشناسی واج در جدول ۴ مشاهده می‌شود. شکل ۸ نمودار مربوط به این روش را نشان می‌دهد.



شکل۸- نمودار کلی تلفیق دانش زبانی با اطلاعات صوتی به روش معکوس‌سازی با اصلاح ورودی به روش گرادیان

۷-۳ تلفیق با استفاده از معکوس سازی شبکه به روش آموزش شبکه معکوس

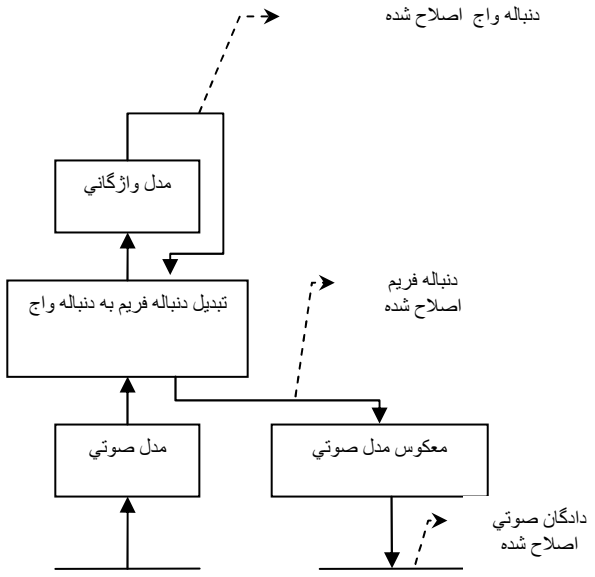
در این روش به علت نیاز به شبکه معکوس مدل صوتی، در هنگام آموزش مدل صوتی مرجع، معکوس این مدل نیز آموزش داده می‌شود(ورودی مدل معکوس دنباله فریم و خروجی آن پارامترهای بازنمایی است).

به این ترتیب مدل معکوس قادر خواهد بود تا از خروجی مدل صوتی، تقریبی از ورودی آن یعنی دادگان صوتی را به‌دست آورد.

مانند روش مطرح شده در بخش قبلی، پس از به‌دست آوردن دنباله واج مدل واژگانی NLPCA با سرپرستی توسط شبکه معکوس، دنباله فریم(خروجی مدل صوتی) مطابق این خروجی اصلاح می‌شود. سپس با استفاده از شبکه معکوس مدل صوتی، دادگان صوتی جدیدی از روی دنباله فریم اصلاح شده به‌دست می‌آید. اعمال پارامترهای بازنمایی به‌دست‌آمده از معکوس مدل صوتی، به مدل صوتی جدیدی که مطابق روش قبلی بر روی دادگان آموزش اصلاح شده، تعلیم دیده است و متوالی‌شدن مدل واژگانی با آن، افزایش قابل ملاحظه‌ای در صحت بازشناسی واج دادگان آزمون اصلاح شده مطابق نتایج جدول۴ از خود نشان می‌دهد. شکل ۹ نمودار کلی طرح را نشان می‌دهد.

جدول۴- نتایج اصلاح واج دادگان آزمون با روش‌های تلفیق دانش زبانی با دادگان صوتی (تشخیص مرز واج از برچسب دادگان)

روش تلفیق	میزان دقت بازشناسی واج	میزان صحت بازشناسی واج	میزان حذف واج	میزان درج واج	میزان جانشینی واج
معکوس سازی با اصلاح ورودی به روش گرادیان	٪۷۹/۴	٪۸۲/۶	٪۷/۸	٪۳/۲	٪۹/۶
معکوس سازی شبکه به روش آموزش شبکه معکوس	٪۷۹/۸	٪۸۲/۸	٪۷/۷	٪۳	٪۹/۵



شکل۹- نمودار کلی تلفیق دانش زبانی با اطلاعات صوتی به روش معکوس‌سازی شبکه به روش آموزش شبکه معکوس

۷-۴ شبکه مرزشناس[۳۸]

در روش‌های تلفیق فوق، در اصلاح دنباله فریم از دنباله واج اصلاح شده، اطلاع از موقعیت مرز واج کاملاً ضروری است. نتایج به‌دست آمده قبلی با در نظرگرفتن مرز واج از برچسب دادگان، محاسبه شده‌اند. به منظور عملی‌کردن روش‌های تلفیق، مرز واج از روی پارامترهای بازنمایی(دادگان صوتی) شناسایی می‌شود. به این منظور در خروجی مدل صوتی یک نورون با تابع غیرخطی سیگموئید اضافه می‌شود. نحوه برچسب‌دهی نورون مرز شناس به این‌گونه است که زمانی‌که مرز بین دو واج دقیقاً در وسط ورودی شبکه قرارگیرد خروجی "۱"، یک فریم جلوتر و عقب‌تر خروجی "۰،۷۵"، دو فریم جلوتر و عقب‌تر خروجی "۰/۲۵" و در غیر از این موارد خروجی "۰" خواهد داشت. درصد صحت بازشناخت مرز دادگان آزمون با درنظر گرفتن دوفریم خطا از سمت چپ و راست مرز ۹۵/۲[۳۹]٪ به‌دست می‌آید.

جدول ۵ نتایج بازشناسی واج حاصل از روش‌های تلفیق، با مرز واج به‌دست آمده از شبکه مرزشناس را نشان می‌دهد.

جدول۵- نتایج اصلاح واج دادگان آزمون با روش‌های تلفیق دانش زبانی با دادگان صوتی(تشخیص مرز واج با شبکه مرزشناس)

روش تلفیق	میزان دقت بازشناسی واج	میزان صحت بازشناسی واج	میزان حذف واج	میزان درج واج	میزان جانشینی واج
معکوس سازی با اصلاح ورودی به روش گرادیان	٪۷۸/۸	٪۸۲/۱	٪۸/۳	٪۳/۳	٪۹/۷
معکوس سازی شبکه به روش آموزش شبکه معکوس	٪۷۹/۲	٪۸۲/۴	٪۸	٪۳/۲	٪۹/۵

۸- بحث و نتیجه گیری

بررسی سامانه درک و بازشناسی در انسان نشان می‌دهد که این سامانه یک ساختار سلسله مراتبی و دوسویه است[۲۹،۳۰،۳۱،۳۲،۳۳]. استفاده از اطلاعات سطوح فوقانی سامانه بازشناخت تصویر در تفسیر و پردازش دادگان ورودی، کارآیی و انعطاف‌پذیری این سامانه را به نحو بسیار مطلوبی افزایش داده است[۲۹]. دوسویه کردن شبکه‌های عصبی جلوسو، افزایش قابل ملاحظه‌ای در کارآیی این شبکه‌ها و تشکیل بسترهای جذب پویا در آن‌ها داشته‌است[۳۴،۳۵،۳۶]. در این تحقیق کارآیی شبکه‌های عصبی دوسویه در بازشناخت گفتار، با ارایه روشی جدید در ترکیب قابلیت‌های شبکه‌های عصبی NLPCA [۲۰] و معکوس‌سازی شبکه‌های عصبی نشان داده شده است. در شبکه‌های NLPCA سؤالی که هنوز پاسخ مشخصی برای آن وجود ندارد این است که آیا مؤلفه‌های اساسی غیرخطی را باید به صورت با سرپرستی و با معلم به شبکه عصبی آموزش داد و یا این که امکان استخراج بهینه و خودکار آن توسط خود شبکه وجود دارد. همان‌گونه که در طرح مدل‌های واژگانی مشاهده‌شد، در مدل NLPCA مؤلفه‌های اساسی به صورت خودکار و توسط شبکه استخراج می‌گردد. در این روش نگاشت فضای ورودی به فضای مؤلفه‌های اساسی غیرخطی و بازسازی مجدد ورودی از این مؤلفه‌ها، امکان فیلترکردن غیرخطی[۴۰] ورودی و تصحیح این دنباله را فراهم می‌سازد. در مدل دوم (مدل NLPCA با سرپرستی) بخشی از فضای مؤلفه‌ها، با سرپرستی و بخشی بدون سرپرستی آموزش می‌بینند. بهبود نتایج حاصل از این روش و نیز مدل سوم حکایت از برتری آموزش با سرپرستی، نسبت به تشکیل مؤلفه‌ها به صورت خودکار دارد. مهم‌ترین مسأله در آموزش با سرپرستی، تعریف صحیح معلم است. لذا در مدل چهارم با ترکیب شبکه معکوس و شبکه NLPCA، تا حدودی این مشکل رفع شده است. با افزایش صحت بازشناسی واج ۹/۴۱٪ حاصل شده از این مدل، به نظر می‌رسد که به‌کارگیری دانش سطوح فوقانی، توفیق بالاتری نسبت به تلاش در جهت بهینه‌سازی مدل صوتی و یا پارامترهای بازنمایی آن خواهد داشت. در ادامه این تحقیق نشان داده شده است که تلفیق دانش واژگانی با پارامترهای بازنمایی نیز می‌تواند مجدداً صحت بازشناسی را افزایش دهد. دوسویه بودن پردازش این امکان را فراهم می‌سازد تا تنوعات و یا عدم انطباق الگوهای ورودی، تا حد ممکن براساس دانش سطوح فوقانی برطرف شود. در این تحقیق دو روش معکوس‌سازی به جهت پردازش دوسویه استفاده شده است. نتایج بهتر روش معکوس‌سازی با استفاده از شبکه معکوس نسبت به اصلاح ورودی به روش گرادیان، به علت گیرافتادن روش‌های مبتنی بر گرادیان، درکمینه‌های موضعی است. از سوی دیگر کاهش صحت بازشناسی در استفاده از شبکه مرز شناس نسبت به تشخیص مرز واج از برچسب دادگان، به دلیل خطای بازشناخت مرز واج است.

در حالی که مدل‌های واژگانی مطرح‌شده برای کلمات مجزا طراحی و پیاده‌سازی شده‌اند؛ امید است در ادامه با طرح روش‌های جدید، نتایج مناسبی نیز با مشخص نبودن مرز کلمات در گفتار پیوسته حاصل شود.

قدردانی

این تحقیق از طرح حمایت مرکز تحقیقات مخابرات ایران از اجرای پروژه‌های دکترا بر طبق قرارداد مورخ ۱۳۸۲/۵/۲۹ و شماره ۵۰۰/۵۴۵۱/ت بهره‌مند گردیده است.

مراجع

[1] L. Lamel and G. Adda, "On Designing Pronunciation Lexicons for Large Vocabulary Continuous Speech Recognition," *Proc. ICSLP96*, pp. 6-9, 1996.

[2] L .Bahl, J. Baker, P. Cohen, F. Jelinek, B. Lewis and R. Mercer, "Recognition of a Continuously Read Natural Corpus," *Proc. ICASSP78*, pp. 422-424, 1978.

[3] T. Fukada and Y. Sagiska, "Automatic Generation of a Pronunciation Dictionary Based on Pronunciation Networks," *Trans. IEICE*, J80-D-II,10, pp. 2626-2635, 2003.

[4] M. Randolph, "A Data-Driven Method for Discovering and Predicting Allophonic Variation," *Proc. ICASSP90*, pp. 1177-1180, 1990.

[5] M. Riley, "A Statistical Model for Generating Pronunciation Networks," *Proc ICASSP91*, pp. 737-740, 1991.

[6] C. Wooters, and A. Stolcke, "Multiple-Pronunciation Lexical Modeling in a Speaker Independent Speech Understanding System," *Proc. ICSLP91*, pp. 1363-1366, 1991.

[7] P. Schntid, R. Cole, and M. Fanty, "Automatically Generated Word Pronunciations from Phoneme Classifier Output," *Proc. ICASSP03*, pp. 223-226, 2003.

[8] T. Sloboda, "Dictionary Learning: Performance Through Consistency," *Proc. ICASSP95*, pp. 453-456, 1995.

[9] T. Imai, A. Ando and E. Miyasaka, "A New Method for Automatic Generation of Speaker-Dependent Phonological Rules," *Proc ICASSP02*, pp. 864-867, 2002.

[10] J. Humphries, *Accent Modeling and Adaptation in Automatic Speech Recognition*, Ph.D. thesis, University of Cambridge, 1997.

[11] E. Fosler, M. Weintraub, S. Wegmann, Y. H. Kao, S. Khudanpur, J. Galles and M. Saraclar, "Automatic Learning of Word Pronunciation from Data," *Proc. ICSLP96*, pp. 28-29,1996.

[12] E. Weintraub, E. Fosler, C. Galles, Y. H. Kao, S. Khudanpur, M. Saraclar and S. Wegmann, "Automatic Learning of Word Pronunciation from Data" *JHU Workshop96 Project Report*, 1996.

[13] B. Byrne, M. Finke, S. Khudanpur, J. McDouough, H. Nock, M. Riley, M. Saraclar, C. Wooters and C. Zavaliagkos, "Pronunciation Modeling for Conversational Speech Recognition: A Status Report front WS97" *Proc. 1997 IEEE Workshop on Speech Recognition and Understanding*, 1997.

[14] T. Sejnowski and C. Rosenberg, *NETtalk: A Parallel Network that Learns to Read Aloud*, The Johns Hopkins Univ. Electrical Engineering and Computer Science Tech. Report JHU/ EECS 86/01. 1986.

[15] M. Ostendorf and H. Singer "HMM Topology Design Using Maximum Likelihood Successive State Splitting," *Computer Speech and Language* 11, pp. 17 41, 1997.

[۱۶] س. ر .معراجی، **بررسی امکان مقاوم سازی مدل شبکه عصبی بازشناخت گفتار نسبت به تغییرات سرعت بیان**، پایان‌نامه کارشناسی ارشد بیوالکتریک، دانشکده مهندسی پزشکی، دانشگاه صنعتی امیرکبیر، ۱۳۸۲ .

[۱۷] الف. نژادقلی، **بازشناخت مقاوم گفتار نسبت به تنوعات مختلف گوینده در شبکه‌های عصبی بازشناخت گفتار**، پایان‌نامه کارشناسی ارشد بیوالکتریک، دانشکده مهندسی پزشکی، دانشگاه صنعتی امیرکبیر، ۱۳۸۲ .

[۱۸] ل. انصاری، **مدل‌سازی اثرات هم‌تولیدی آواها در یک مدل شبکه عصبی بازشناخت گفتار**، پایان‌نامه کارشناسی ارشد بیوالکتریک، دانشکده مهندسی پزشکی، دانشگاه صنعتی امیرکبیر، ۱۳۸۲ .

[۱۹] ک. کریمی، **به‌کارگیری مشخه‌های گوینده در جهت بهبود کیفیت مدل‌های بازشناخت گفتار**، پایان‌نامه کارشناسی ارشد بیوالکتریک، دانشکده مهندسی پزشکی، دانشگاه صنعتی امیرکبیر، ۱۳۸۱.

[20] K. I. Diamantaras, *Neural Networks and Principal Component Analysis* , Handbook of Neural Network Signal Processing , CRC Press, 2002.

[21] M. A. Kramer, "Nonlinear Principal Component Analysis Using Autoassociative Neural Networks," *AICIE Journal*, vol. 37, no. 2, pp. 233-243, 1991.

[22] C. A. Jensen, R. D. Reed, R. J. Marks and et.al, "Inversion of Neural Networks: Algorithms and Applications," *Proc. IEEE, Neural Networks*, vol. 87, no. 9, 1999.

[23] R. J. Williams, "Inverting a Connectionist Network Mapping by Backpropagation of Error", *Proc. 8th Annu. Conf. Cognitive Science Society*, pp. 859-865, 1986.

[۲۴] س. ع. سیدصالحی، **افزایش کارآیی بازشناخت الگوی شبکه‌های عصبی جلوسو از طریق توسعه روش‌هایی برای دو سویه کردن عملکرد آن‌ها**، گزارش طرح پژوهشی، دانشکده مهندسی پزشکی، دانشگاه صنعتی امیرکبیر، ۱۳۸۳.

[25] M. Bijankhan, et. al, "FARSDAT-the Speech Database of Farsi Spoken Language," *SST94*, pp. 826-831, 1994.

[۲۶] م. رحیمی‌نژاد، **توسعه و بهبود کیفیت روش‌های استخراج پارامترهای بازنمایی در سیستم‌های بازشناخت گفتار**، پایان‌نامه کارشناسی ارشد بیوالکتریک، دانشکده مهندسی پزشکی، دانشگاه صنعتی امیرکبیر، ۱۳۸۱.

[۲۷] م. ولی، و س. ع. سیدصالحی، «ارزیابی کارآیی دو بازنمایی MFCC و LHCB در بازشناسی مقاوم به تنوعات گفتار مستقیم و تلفنی»، **مجموعه مقالات دهمین کنفرانس سالانه انجمن کامپیوتر ایران**، ص۳۰۵-۳۱۲، ۱۳۸۳.

[28] D. Nguyen and B. Widrow, "Neural Networks for SelfLearning Control Systems," *IEEE Control Systems Magazine*, pp. 18-23, 1990.

[29] E. Koerner, M. O. Gewaltig, U. Koerner, A. Richter and T. Rodemann, "A Model of Computation in Neocortical Architecture," *Neural Networks Elsevier Science*, vol. 12, pp. 989-1005,1999.

[30] E. Koerner, H. Tsujino and T. Masutani , "A Cortical-Type Modular Neural Network for Hypothetical Reasoning," *Neural Networks , Elsevier Science*, vol. 10 , no. 5 , pp. 791-814, 1997.

[31] E. Koerner and G. Matsumoto, "Cortical Architecture and Self-referential Control for Brain- Like Computation, a New Approach to Understanding How the Brain Organizes Computation," *IEEE Eng. In Medicine and Biology Magazine*, 2002.

[32] J. Ghosen and Y. Bengio, "Bias Learning, Knowledge sharing," *IEEE Trans. On Neural Networks*, vol. 14, no. 4, 2003.

[33] M. M. Mesulam, *From Sensation to Cognition*, Brain, Oxford Univ. Press, pp. 1013- 1052, 1998.

[34] L. K. Saul and M. I. Jordan, "Attractor Dynamics in Feedforward Neural Networks," *Neural Computation, Massachusetts Institute of Technology*, vol. 12 , pp. 1313-1335 , 2000.

[35] T. P. Trappenberg, "Continuous Attractor Neural Network," *in Recent developments in biologically*

inspired computing , L. N. , Castro, and F. J. , Von Zuben (eds.) , 2003.

[36] Y. Wu and D. A. Pados , "A Feedforward Bidirectional Associative Memory" , *IEEE Trans. On Neural Networks*, vol. 11, no. 4, 2000.

[1]Pronunciation Variation
[2] Variation in Speaking Style
[3] Out of Vocabulary (OOV)
[4] Pronunciation Dictionary
[5] Large Vocabulary Continuous Speech Recognition(LVCSR)
[6] Phonological Rules
[7] Pronunciation Modeling
[8] Pronunciation Network
[9] Canonical Pronunciation
[10] Alternative Pronunciation
[11] Nonlinear Principal Component Analysis (NLPCA)
[12] Principal Component Analysis(PCA)
[13] Kramer
[14] Bidirectional
[15] Actual Telephone Database(TFARSDAT Database)
[16] Bark Scale
[17] Mel Scale
[18] Mel Frequency Cepstral Coefficients
[19] Normalization
[20] Acoustic Model
[21] Hyperbolic Tangent
[22] Nguyen-Widrow
[23] Resilient Backpropagation Algorithm
[24] Correction Ratio
[25] Accuracy Ratio
[26] Deletion Ratio
[27] Insertion Ratio
[28] Substitution Ratio
[29] کلیه نتایج با نرم‌افزار NIST محاسبه شده است.
[30] Variability
[31] Mismatch
[32] Attractor
[33] Basin of Attraction
[34] $:بست چ، 'چ، /:آ، =:بست د است.
[35] Isolated Word
[36] Word Code
[37] Pronunciation Transcript Code
[38] Border Detector
[39] به دلیل برچسب‌دهی دادگان به صورت دستی توسط اپراتور، تا دو فریم از چپ و راست مرز را نمی‌توان خطای بازشناخت مرز در نظر گرفت.
[40] linear Filtering

محمدرضا یزدچی کارشناسی خود را در مهندسی برق- الکترونیک از دانشگاه صنعتی اصفهان در سال ۱۳۷۶ و کارشناسی ارشد را از دانشگاه صنعتی امیرکبیر در مهندسی پزشکی- بیوالکتریک در سال ۱۳۷۸ دریافت نموده است. او در حال حاضر دانشجوی دوره دکتری مهندسی پزشکی– بیوالکتریک در دانشگاه صنعتی امیرکبیر می‌باشد. زمینه‌های تخصصی وی پردازش سیگنال‌های حیاتی، پردازش گفتار و شبکه‌های عصبی مصنوعی و زیستی می‌باشد.

سیدعلی سیدصالحی کارشناسی خود را در مهندسی برق از دانشگاه صنعتی شریف در سال ۱۳۶۱، کارشناسی ارشد را در مهندسی برق از دانشگاه صنعتی امیرکبیر در سال ۱۳۶۷ و دکتری خود را در مهندسی برق-بیوالکتریک از دانشگاه تربیت مدرس در سال ۱۳۷۴ دریافت نموده است. وی درحال حاضر استادیار دانشکده مهندسی پزشکی دانشگاه صنعتی امیرکبیر می‌باشد. زمینه‌های پژوهشی مورد علاقه او پردازش و بازشناسی گفتار، شبکه‌های عصبی مصنوعی زیستی ، مدل‌سازی و عملکرد مغز و پردازش خطی و غیرخطی سیگنال می‌باشد.

فرشاد الماس‌گنج در سال ۱۳۶۳ در رشته برق گرایش الکترونیک در مقطع کارشناسی از دانشگاه صنعتی امیرکبیر فارغ‌التحصیل گردیده است. سپس در همین گرایش دوره کارشناسی ارشد خود را در سال ۱۳۶۷ به پایان رسانید و به سمت عضو هیأت علمی دانشگاه صنعتی امیرکبیر مشغول به‌کار گردید. بعد از تاخیری کوتاه، تحصیل خود را در رشته برق، گرایش مهندسی پزشکی در دانشگاه تربیت مدرس ادامه داد. او در سال ۱۳۷۷ به درجه دکتری دست یافت. از آن زمان تا کنون در دانشکده مهندسی پزشکی دانشگاه صنعتی امیرکبیر با سمت استادیاری حضور دارد. زمینه تخصصی مورد علاقه او پردازش سیگنال و خصوصاً پردازش انواع سیگنال‌های گفتاری است.