بازشناسی گفتار توسط شبکه‌های عصبی مدولار بر مبنای بکارگیری توأم نواحی صوتی یکنواخت و گذرا

سید علی سید صالحی

استادیار

دانشکده مهندسی پزشکی، دانشگاه صنعتی امیرکبیر

چکیده

کارهای قبلی بر روی بازشناسی گفتار به کمک شبکه‌های عصبی غالباً بر تقطیع واجی و یا نگاشت مسیر بردار بازنمایی گفتار به فضای واج‌ها متکی بوده‌اند. در این روش‌ها به علت نحوهٔ برچسب‌دهی در مرزهای واج‌ها بخشی از اطلاعات مفید گفتار ممکن است از دست بروند. برخی نتایج اخیر نشان می‌دهند که اولاً نواحی مرزی و گذرای واج‌ها برای شناسائی به عنوان واحدهای صوتی پتانسیل خوبی دارند و دوم اینکه بازشناسی نواحی گذرای سریع آواها به عنوان نشانه‌های صوتی ثابت توسط شبکه‌های عصبی به کیفیت بازشناسی بالائی منجر گردیده است. بازشناسی مجزای نواحی یکنواخت و نواحی گذرای آواها بر مبنای تفکیک همخوان، واکه و ترکیب نتایج آنها را بر روی گفتار فارسی پیاده‌سازی نموده‌ایم. بررسی خطاهای عملکرد این مدل نشان داده‌اند که برخی نکات نظری زبان‌شناختی پیرامون همخوان‌ها و واکه‌ها با نتایج پیاده‌سازی‌های مهندسی کاملاً تطبیق نمی‌کنند.

کلمات کلیدی

بازشناسی گفتار، شبکه‌های عصبی مدولار، نواحی صوتی یکنواخت و گذرا، واکه ـ همخوان، بازشناسی واج.

A Modular Neural Network Speech Recognizer Based on the Both Acoustic Steady Portions and Transitions

S. A. Seyyed Salehi
Asistant Professor
Biomedical Engineering Department,
Amirkabir University of Technology

Abstract

Previous works on speech recognition utilizing neural networks have often relied on either recognition through segmentation or mapping of the representation trajectories to the phoneme space. Here, information could be missed due to the method of border labeling techniques.

Recent works have indicated that firstly, phonetic borders and transitions would have a good potential to be recognized as acoustic units, and secondly, recognition of the fast transitions by neural networks, as fixed cues in time, results in high performance detection and recognition of those events. This approach was manifested through recognition of basic units formed from the VC and CV borders in Farsi (Persian) spoken language. Analysis of the resulting errors has indicated certain discrepancies amongst the theoretical linguistic points of view and implementation outcome.

Keywords

Speech Recognition, Modular Neural Networks, steady & Transition, Vowel-Consonant, Phoneme Recognition.

مقدمه

بررسی نتایج تحقیقات قبلی در بازشناسی واج‌های[1] گفتار و تجزیه و تحلیل خطاهای موجود در بازشناسی واج‌ها در روش‌های پیاده‌سازی شده پیشین نشان داده‌اند که اطلاعات صوتی و گفتاری مفید و قابل ملاحظه ای را از نواحی گذرای مرزی آواها می‌توان استخراج نمود.

مدل‌های شبکه عصبی پیاده‌سازی شدهٔ قبلی برای بازشناسی واج‌ها، غالباً یا متکی به بازشناسی واج‌ها پس از تقطیع[2] بوده‌اند [۱] و یا اینکه بر مبنای نگاشت کل مسیر بردار بازنمائی به فضای واج‌ها عمل می‌کردند [۲، ۳، ۴]. در هر حال دقت در نتایج عملکرد این مدل‌ها نشان می‌دهد که در این روش‌ها به دلیل نحوه برچسب‌دهی در نواحی مرزی آواها، اطلاعات صوتی موجود و نهفته در این نواحی عمدتاً مورد کم‌توجهی قرار می‌گرفته‌اند.

در هر دو روش ذکر شدهٔ فوق برای بازشناسی واج‌ها، نواحی با مشخصات طیفی سریعاً گذرای سیگنال گفتار درتقطیع و یا برچسب‌دهی در مرز ما بین واج‌ها قرار می‌گیرند و با توجه به وجود خطاهای متعددی نظیر خطا در تشخیص محل تقطیع یا برچسب‌دهی توسط انسان و نیز خطای در انطباق زمانی مدل بازشناسی با سیگنال صوتی در هنگام تعلیم و بازشناسی، ایجاد تنوعات و چندگونگی‌های زائدی می‌نماید که موجب اتلاف و ریزش اطلاعات صوتی مفید موجود در سیگنال گفتار در هنگام بازشناسی می‌گردد. این نکته می‌تواند به عنوان یکی از تنگناهای موجود در مسیر عملکرد سیستم‌های بازشناسی گفتار مورد توجه و ملاحظه قرار گیرد. برای رفع این مشکل لازم است که روش‌های رایج برچسب‌دهی خروجی مدل‌های بازشناسی واحدهای صوتی را تغییرداده و در خود واحدهای صوتی پایه نیز اصلاحاتی را به عمل آوریم.

در تعلیم واج‌ها به شبکه عصبی به صورت تقطیع شده، تقطیع آواها در نواحی‌ای از سیگنال گفتار که بیشترین سرعت تغییرات طیفی را دارند، یعنی مرزآواها انجام می‌شود. در چنین حالتی اولاً کوچکترین خطا در تشخیص مدل تقطیع، بیشترین تأثیرات نامطلوب را درتعلیم و بازشناسی مدل ایجاد می‌نماید و حساسیت پارامترهای طیفی به تغییرات محل تقطیع در این نواحی بیشینه[3] است. این در حالی است که نواحی گذرا درتعلیم به مدل می‌بایست دارای کمترین مقدار خطا باشند. درتعلیم نگاشت کل مسیر بردار بازنمایی به فضای واجها به شبکه عصبی، نیز گرچه سیگنال گفتار تقطیع نمی‌شود [۲، ۳، ۴] لیکن بازهم تغییر برچسب‌ها در خروجی شبکه در نواحی گذرای مرزی آواها، اتفاق می‌افتد که هم دارای خطای تشخیص انسانی ناشی از برچسب‌دهی دستی دادگان است و هم اینکه بطوریکه عملاً نتایج پیاده‌سازی‌های قبلی نشان داده‌اند بیشترین خطای بازشناسی را در نواحی مرزی آواها داریم. بنابراین اگر ما تغییراتی در نحوه بیان برچسب‌های سیگنال گفتار در هنگام تعلیم به شبکه بدهیم به نحوی که نقاط ضعف ذکر شده در بالا در آنها به حد کمینه[4] رسیده باشند می‌توان انتظار داشت که به این وسیله به صحت بازشناسی بالاتری بتوان دست یافت [۵، ۶، ۷].

نکته دیگری که در این رابطه نیاز است مورد توجه قرار گیرد این است که در سیگنال گفتار، نواحی گذرا نسبت به نواحی یکنواخت خصوصاً واکه‌ها، کمتر تحت تأثیر سرعت بیان قرار می‌گیرند و مشخصات دینامیکی پایدارتری دارند و به بیان دیگر اینکه طول‌های زمانی وقایع گذرای گفتار در عبارات مختلف، نسبت به نواحی ایستان گفتار دارای تغییرات کمتری هستند و افراد در هنگام تغییر سرعت بیان خود نوعاً بیشتر طول نواحی ایستان خصوصاً واکه‌های گفتار خود را تغییر می‌دهند [۸] و این برای تعلیم نواحی گذرا به مدل شبکه عصبی یک مزیت نسبی به حساب می‌آید. زیرا تغییرات و جابجائی در مشخصات زمانی وقایع صوتی در هنگام تعلیم و بازشناسی توسط شبکه‌های عصبی، تأثیرات نامطلوبی در کیفیت عملکرد شبکه‌های عصبی باقی می‌گذارد. بنابراین اگر ما دسته‌ای از واحدهای صوتی مورد بازشناسی را در نواحی گذرای گفتار اختیار کنیم می‌توانیم با کمک آنها نقاط ضعف ذکر شده در بالا را تا حدی برطرف نمائیم.

۱ـ اصول و مبانی، ساختار و اجزاء مدل بازشناسی گفتار(نمودار۱)

مشخصات طیفی سیگنال گفتار متغیربا زمان است و وقایع صوتی درون آن متأثر از رفتار اجزای تولیدی گفتار[5] در ناحیه صوتی است. هرتغییر موقعیت هر یک از این عناصر تولید گفتار، حداقل یک تغییر مشخصات طیفی را در سیگنال گفتار در پی دارد. دقت در نحوه ادراک گفتار توسط انسان نشان می‌دهد که این تغییرات طیفی در سیگنال گفتار دارای هویت و بار اطلاعاتی قابل ملاحظه‌ای بوده و به عنوان مشخصه‌ها و وقایع اصلی در سیگنال گفتار نقش ایفا می‌کنند. در رویکرد مطرح در

این مقاله، ما نواحی گذرای مابین دو واج مجاور با توجه به واکه یا همخوان بودن آنها یعنی مرزهای CV، VC و CC را به عنوان واحدهای کمکی جهت بازشناسی درنظر گرفته‌ایم. همچنین در بررسی نتایج پیاده‌سازی مدل بازشناسی واج‌ها به کمک مشخصه‌های تولیدی که قبلاً انجام داده‌ایم، [۲، ۴] مشاهده گردید که مشخصه تولیدی هجائی[۶] که بیانگر واکه یا همخوان بودن یک واج است توسط شبکه عصبی با دقت بالائی قابل یاد گرفتن است و میزان صحت بازشناسی بالائی را نیز دارد. از سوی دیگر واکه‌ها با صحت بالائی توسط شبکه‌های عصبی شناسائی می‌شوند و خطاهای بازشناسی عمدتاً مربوط به همخوان‌هاست. براین اساس در این جا مدل بازشناسی را به این صورت طراحی می‌کنیم که ابتدا توسط یک شبکه عصبی، نواحی واکه و همخوان شناسائی می‌شوند (VCS) و علاوه بر این به منظور تعیین دقیق موقعیت گذرای مابین واکه‌ها و همخوانها یک شبکه عصبی جداگانه توسط سیگنال طوری تعلیم داده می‌شود که برای هر مرز بین دو واج مجاور، یک خروجی شبکه بصورت لحظه‌ای فعال شود و یک پالس باریک مشابه ضربه در خروجی مربوطه بدهد. آزمایشات مقدماتی نشان دادند که شبکه عصبی توانائی خوبی در یادگیری وقایع صوتی بصورت لحظه‌ای (ضربه‌ای)[۷] و ثابت با زمان دارد. هنگامیکه در این روش از شبکه عصبی برای بازشناسی وقایع لحظه‌ای صوتی استفاده می‌کنیم می‌توانیم برای جبران خطای برچسب‌دهی دادگان، از روش بهترین انطباق نیز استفاده کنیم و شبکه عصبی را هنگام تعلیم و بازشناسی وقایع ثابت صوتی ابتدا از نظر زمانی با سیگنال منطبق نمائیم. با این عمل تنوعات موجود در الگوهای ورودی کاهش یافته و کار بازشناسی توسط شبکه عصبی تسهیل می‌گردد. تحقیقات زبانشناختی، ساختارهای CVCC, CVC, CV را برای هجاهای فارسی مطرح کرده‌اند [۹]. لذا ما ابتدا موقعیت واکه‌ها را به عنوان قلب و مرکز هجا در گفتار مشخص کرده و سپس نوع واکه را مورد شناسائی قرار می‌دهیم. از سوی دیگر به منظور تعیین دقیق نواحی گذرای مرزی مابین آواها، لحظات گذر ما بین آواها را به یک شبکه عصبی با سه خروجی به نحوی تعلیم می‌دهیم که هر خروجی، نوع مرز یعنی CV، VC و CC را نیز مشخص نماید. در آغاز هر واکه بایستی یک خروجی CV بطور لحظه‌ای فعال گردد و در پایان هر واکه نیز خروجی VC باید فعال شود. علت تعیین جداگانه مرز واج‌ها توسط این شبکه بدست آوردن دقت بالا در تعیین زمان گذر مابین دو واج مجاور است. در برچسب‌دهی خروجی این شبکه بصورت لحظه‌ای برای تعلیم، از یک تابع تقریباً به شکل ضربه به عرض سه فریم و با مقادیر خروجی به ترتیب ۰/۳، ۱ و ۰/۳ در خروجی شبکه استفاده می‌کنیم. برای هر مرز دو واج فقط خروجی مربوط به آن مرز یعنی فقط یکی از حالات VC، CV و CC به صورت ضربه فعال می‌شود. بیان خروجی شبکه بصورت ضربه، علاوه بر تعیین دقیق موقعیت زمانی مرز، بار یادگیری تنوعات صوتی در مرزهای واج‌ها را برای شبکه به حداقل می‌رساند و کار شبکه را آسانتر می‌سازد.

پس از این مرحله، توسط شبکه‌های عصبی جداگانه‌ای نوع همخوان‌های موجود در قبل از واکه (CV) و بعد از واکه (VC) را شناسائی می‌کنیم. در این شناسائی، تعلیم و بازشناسی همخوان ابتدا بصورت ثابت و با توجه به برچسب دادگان انجام می‌شود و سپس به شبکه اجازه داده می‌شود که با استفاده از تکنیک بهترین انطباق، از نظر زمانی ورودی خود را با مرز بین دو واج در سیگنال گفتار منطبق نماید و بدینوسیله خطاهای موجود در تشخیص انسانی موقعیت زمانی مرز بین واج‌ها در برچسب‌دهی دادگان را به حداقل برساند و دادگان تعلیم و تست را برای شبکه عصبی بازشناس واضح و شفاف نماید.

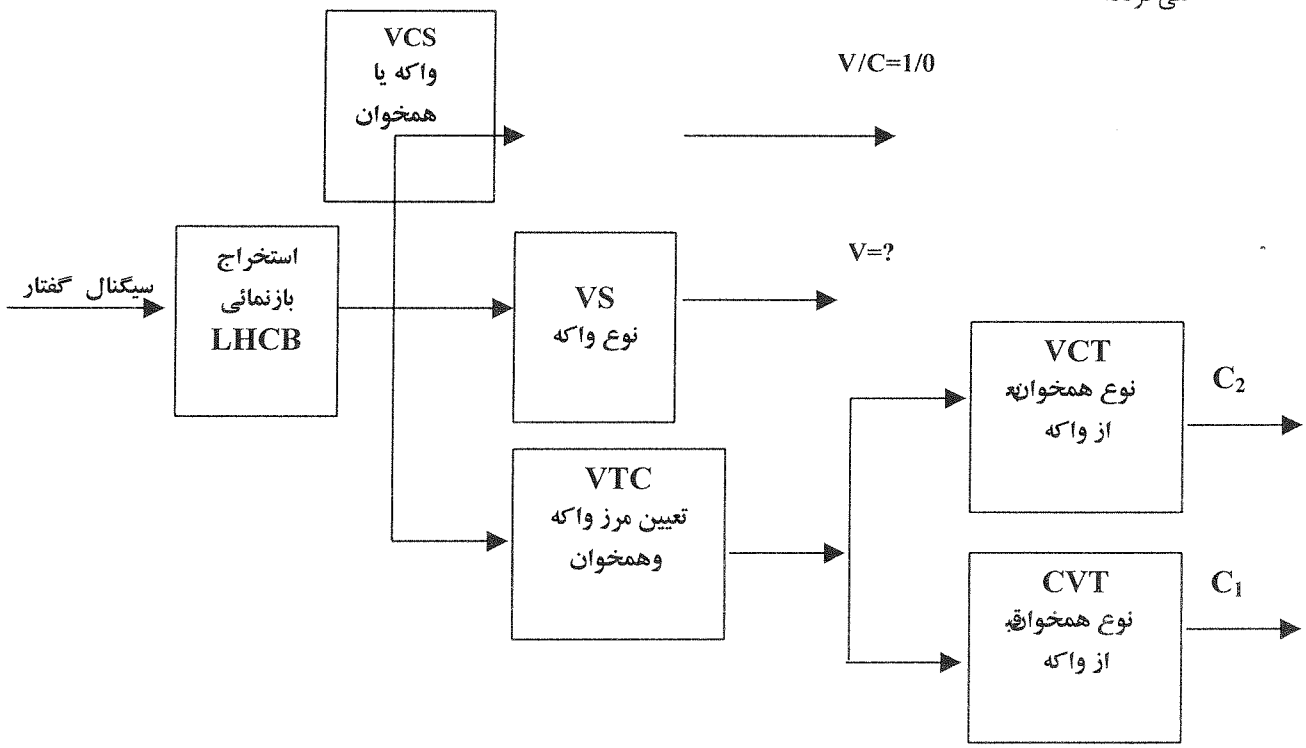
در این مقاله از شناسائی همخوان سوم در هجاهای نوع CVCC که در گفتار محاوره‌ای فارسی کمتر وجود دارد و نقش آن در تشخیص کلمات هم خیلی تعیین کننده نیست فعلاً صرفنظر کرده‌ایم و عملاً هجاهای فارسی را محدود به دو نوع CV و CVC درنظر گرفته‌ایم.

۱-۱ـ بازنمائی گفتار

به عنوان بازنمائی در این مقاله ما از پارامترهای LHCB استفاده کرده‌ایم [۲، ۱۰]. بردار بازنمائی حاوی ۱۸ پارامتر است که هر یک لگاریتم انرژی درون یکی از فیلترهای بانک فیلترهای با باندهای بحرانی (CBFB) به شکل مجذور هنینگ به عرض یک در مقیاس فرکانسی بارک[۸] می‌باشد. در برخی شبکه‌ها از ۱۸ پارامتر تفاضلی بازنمائی LHCB نیز استفاده کرده‌ایم. عرض پنجره زمانی ۱۰۲۴، گام پیشروی فریمها ۵۱۲ و فرکانس نمونه‌برداری ۴۴۱۰۰ نمونه در ثانیه است و دادگان مورد استفاده فارس دات [۱۱] است.

به منظور نرمالیزه کردن پارامترهای بازنمائی نسبت به بلندی صدا و فاصلهٔ میکروفون ما ابتدا فریم‌های گفتاری را برحسب

انرژی مرتب می‌کنیم و ۱۰ درصد بالای پرانرژی‌ترین فریمها را اختیار کرده و میانگین انرژی آنها را محاسبه می‌کنیم و این کمیت را مبنای نرمالیزه و ثابت اختیار می‌کنیم. این به معنی آن است که معیار بلندی صدا در صحبت بیشتر واکه‌های بلند هستند و لذا آنها مبنا قرار داده می‌شوند و از طرفی انرژی موضعی و نسبی فریمها هم که حاوی اطلاعات مفید است حفظ می‌گردد.



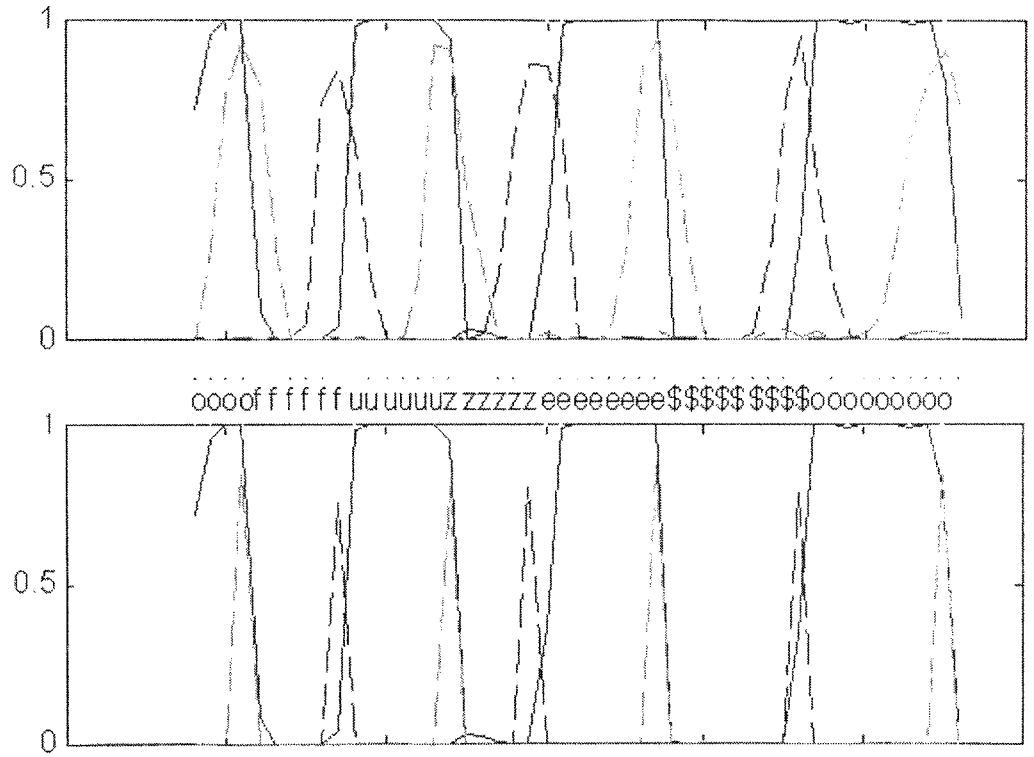
نمودار (۱) نمودار بلوکی و خروجی‌های مدل، VCS: لغزنده (تشخیص واکه یا همخوان بودن)، VS: ثابت یا لغزنده (تشخیص نوع واکه)، VTC: لغزنده (تشخیص محل و نوع مرز واکه همخوان)، VCT و CVT: ثابت و با بهترین انطباق (تشخیص نوع همخوان در قبل و بعد از واکه).

۱-۲ـ شبکه تمیزدهنده واکه ـ همخوان (VCS)[۹]

شبکه عصبی VCS با ساختار مشابه TDNN و با مقدار گره‌های لایه‌ها به ترتیب ۳۷ در ورودی۷×۳۲ در لایه پنهان اول (۷ تعداد تأخیر زمانی)، ۶۴ در لایه پنهان دوم و ۱ در خروجی به منظور تمایز واکه از همخوان بکار رفته است. این شبکه روی مسیر زنجیرهٔ فریم‌های گفتاری رو به جلو حرکت می‌کند ودر هر زمان مشخص می‌کند که ورودی شبکه در یک واکه (۱ در خروجی) و یا همخوان (صفر در خروجی) قرار گرفته است. در اینجا به عنوان ورودی؛ خود بردار بازنمائی، تفاضل دو بردار بازنمائی متوالی و انرژی مورد استفاده قرار گرفته‌اند. در این شبکه ۷ فریم متوالی همزمان به شبکه داده می‌شوند که در لایه دوم باهم ترکیب می‌شوند. پس از تعلیم این شبکه با دادگان تعلیم، از آن برای تشخیص واکه و یا همخوان در مدل بازشناسی استفاده کرده‌ایم. خروجی این شبکه عمدتاً برای تجزیه و تحلیل عملکرد مدل از طریق مشاهده موقعیت هر فریم نسبت به واکه یا همخوان بودن آن و تفکیک چشمی قسمتهای مختلف هجاها و واجها و خصوصاً در تحلیل علت خطاهای موجود در عملکرد مدل، مورد استفاده قرار می‌گیرد (نمودار ۲).

۱-۳ـ شبکه تمیزدهنده مکان و نوع مرز بین واج‌ها (VTC)

شبکه عصبی VTC با ابعاد ۳۷ گره در ورودی، ۳۲×۱۰ نورون در لایه پنهان اول، ۱۲۸ نورون در لایه پنهان دوم و ۳ نورون در خروجی برای طبقه‌بندی نوع مرزها (CC, CV, VC) بکار رفته است. ۳۷ ورودی در اینجا مشابه شبکه قبل است و ۱۰ فریم متوالی همزمان به شبکه داده می‌شوند (۵ فریم در هر طرف مرز) تا شبکه توان شناسائی دینامیکها را نیز داشته باشد.



نمودار(۲) نمونه عملکرد موفق شبکه VCS (خط پر)وتطابق آن با خروجیهای CV وVC شبکهVTC (خط چین).

برچسب‌دهی خروجی بصورت ضربهٔ باریک می‌باشد که قله ضربه مرز را مشخص می‌کند. برای این منظور ما از مقادیر متوالی (...،۰/۳،۰،۰/۳،۱،...) در هنگام تعلیم شبکه استفاده کرده‌ایم. مقدار ۱ هنگامی در برچسب خروجی مطلوب شبکه داده می‌شود که موقعیت مرز در وسط ده فریم ورودی شبکه قرار می‌گیرد. در نواحی غیرمرزی خروجیها همه صفرند و در مرزها، مقادیر غیرصفر فقط به آن خروجی مربوط به آن مرز داده می‌شود. این نحوه برچسب‌دهی، شبکه را قادر می‌سازد تا وقایع گذرای صوتی را بصورت ثابت و لحظه‌ای شناسائی کند و موقعیت آنها را نیز تشخیص دهد. علاوه بر این، این امکان به ما داده می‌شود که هنگام تعلیم و تست شبکه با دادگان از روش بهترین انطباق (Best-Matched, B.M.) استفاده کنیم. در بکارگیری این روش B. M. خروجی شبکه برای یک تعداد فریم متوالی بدست آورده می‌شود و سپس این خروجیها با مقادیر برچسب دادگان مربوط به این فریمها از نظر شباهت مورد بررسی قرار می‌گیرند. با لغزاندن این دو الگوی خروجی شبکه و خروجی مطلوب بر رویهم و پیدا کردن بهترین موقعیت انطباق آنها، آنگاه شبکه را با الگوی برچسب خروجی که بهترین انطباق را یافته تعلیم می‌دهیم. با این روش ما اثر خطای موجود در برچسب‌دهی انسانی دادگان را تا حدی برطرف می‌کنیم و به شبکه این امکان را می‌دهیم که با دادگان واضح‌تر و شفاف‌تری تعلیم ببیند. به این ترتیب پس از تعلیم این شبکه، هنگام بازشناسی مرزها این شبکه بر روی فریم‌های متوالی به جلو حرکت می‌کند و برای هر مرز بین دو واج یکی از خروجی‌هایش فعال می‌شود و ضمن تعیین موقعیت مرزها، نوع مرز (CC, CV, VC) نیز توسط خروجی تعیین می‌گردد. ما به این تعیین نوع مرز از آن جهت نیاز داریم که بتوانیم نوع شبکه مربوط به آنرا برای تشخیص همخوان مجاور مرز مورد استفاده قرار دهیم.

با داشتن خروجی‌های این شبکه عملاً ساختار هجائی گفتار در اختیار ما قرار می‌گیرد (CVCC, CVC, CV) و برای انواع هجاها به صورت دنباله‌ای متشکل از واکه‌ها (V) و همخوان‌ها (C) قابل نمایش می‌باشد. به کمک دو شبکه فوق عملاً ما می‌توانیم سیگنال گفتار را تقطیع واجی و هجائی نمائیم و به بیان دیگر موقعیت مکانی انواع وقایع صوتی موردنظر در اینجا (CC, CV, VC) را مشخص نمائیم. به نظر ما این نحوه برخورد در بازشناسی گفتار با واقعیت غیرخطی سیگنال گفتار تناسب بیشتری دارد. با این شیوه طراحی مدل شبکه عصبی ما توان آن را برای تحمل سرعت‌های مختلف بیان گفتار افزایش می‌دهیم و به مدل نسبت به تغییرات طول واج‌های پیوسته قدرت انعطاف بیشتری می‌دهیم.

۱-٤ـ شبکه واکه‌شناس (VS)

پس از مشخص شدن موقعیت واکه‌ها و همخوان‌ها می‌توانیم شناسائی نوع واکه‌ها و همخوان‌ها را انجام دهیم. به منظور تشخیص نوع واکه به دو راه می‌شود عمل‌کرد یکی اینکه یک شبکه عصبی بصورت ثابت روی واکه قرار گیرد و آنرا شناسائی نماید و دیگر اینکه مشابه روشی که برای تشخیص واکه یا همخوان بودن گفتار بکار گرفتیم، از یک شبکه عصبی که بر روی مسیر فریمهای گفتار به جلو می‌لغزد و برحسب اینکه روی چه واکه‌ای قرار دارد خروجی مربوط به آن یک می‌شود، استفاده کنیم. در اینجا ما از روش اول استفاده می‌کنیم (شبکه VS). ابعاد این شبکه به ترتیب: ۳۷ گره در ورودی، ۳۲×۱۰ نورون در لایه پنهان اول (۳۲ نورون برای هر فریم و ۱۰ فریم متوالی به شبکه داده می‌شوند)، ۱۲۸ نورون در لایه پنهان دوم و ۶ نورون به تعداد واکه‌ها در خروجی، می‌باشند. خروجی شبکه نوع واکه را مشخص می‌کند و فریم اول ورودی این شبکه روی فریم اول واکه قرار داده می‌شود. با داشتن نوع واکه، کار بعدی ما این است که همخوانها را شناسائی کنیم.

در هجاهای فارسی ما سه نوع مدل هجائی داریم CVC،CV و CVCC. بر این اساس یک همخوان در یکی از موقعیت‌های C_1 و C_2 و C_3 در این سه مدل هجائی می‌تواند قرار بگیرد. قبل از واکه C_1V در هر سه مدل، بعد از واکه در مدل دوم و سوم C_1VC_2 و به عنوان همخوان سوم در مدل هجائی $C_1VC_2C_3$. ما این همخوان‌ها را توسط شبکه‌های ثابت جداگانه مورد تعلیم و شناسائی قرار می‌دهیم.

۱-۵ـ شبکه بازشناس نوع همخوان در موقعیت قبل از واکه (CVT)

برای این هدف از یک شبکه عصبی به ابعاد ۱۹ گره در ورودی ۱۱×۳۲ نورون در لایه پنهان اول، ۱۲۸ نورون در لایه پنهان دوم و ۲۳ نورون در خروجی استفاده می‌کنیم. تعداد ۲۳ نورون در خروجی به تعداد ۲۳ همخوان زبان فارسی انتخاب شده که در وضعیت CV می‌توانند بیایند. تعداد فریم اختیار شده در دو طرف مرز CV به تعداد ۷ فریم در سمت همخوان و ۴ فریم در سمت واکه می‌باشد. این شبکه بر روی مرزهای CV بصورت ثابت شده، طوری تعلیم داده می‌شود که بتواند در خروجی نوع همخوان قبل از واکه را شناسائی نماید. در تعلیم و بازشناسی این شبکه نیز از روش بهترین انطباق استفاده می‌کنیم. در اجرای روش بهترین انطباق برای تعلیم و تست باید توجه داشت که گاهی ممکن است دو مرز خیلی به هم نزدیک باشند و این موضوع را در برچسب‌دهی باید اعمال نمود. برچسب‌دهی در خروجی این شبکه هم بصورت ضربهٔ باریک و لحظه‌ای می‌باشد. وقتیکه این شبکه در روی یک مرز CV قرار می‌گیرد بسته به اینکه نوع همخوان چه باشد یک خروجی از خروجی‌های آن که مربوط به آن همخوان است بصورت لحظه‌ای فعال می‌گردد و بقیه غیرفعال باقی می‌مانند. این شبکه وظیفه دارد که در روی مرزهای CV بصورت یک ضربه باریک در خروجی، نوع همخوان را شناسائی نماید. یک نکته که در دل این نحوه شناسائی همخوان نهفته است این است که در خیلی موارد اطلاعات مفیدی مرتبط با نوع همخوانها در سیگنال صوتی، در نواحی گذرای سیگنال هنگام ورود و خروج به آن همخوان، وجود دارد خصوصاً در مواردی که واج مجاور یک واکه باشد. با این روش خواسته‌ایم که این اطلاعات را از سیگنال صوتی استخراج نمائیم. نتایج پیاده‌سازی این شبکه نکات جالبی را دربردارد. هنگامیکه موقعیت مرز واجها بدون خطا تشخیص داده شده باشد یعنی از خود برچسب‌های دادگان برای تعیین محل مرز CV در تعلیم و تست استفاده می‌کنیم، میزان صحت بازشناسی همخوان‌ها در حالت مستقل از گوینده بسیار بالا یعنی برابر ۸۵٪ است.

همچنین درصد مواردیکه همخوان صحیح در یکی از دو انتخاب اول یا دوم باشد ۹۱/۲٪ ، درصد مواردیکه همخوان صحیح در یکی از سه انتخاب اول تا سوم باشد ۹۴/۵٪ و در یکی از چهار انتخاب اول برابر ۹۶/۷٪ است.

این نتایج دو نکته را دربر دارد. اول اینکه نشان می‌دهد منشأ شناسائی غلط همخوان‌ها در روش‌های متعارف، عمدتاً به تشخیص غلط مرزهای CV و VC برمی‌گردد که یا محل آنها درست شناسائی نمی‌شود و یا اینکه مرز حذف شده و یا اینکه مرزهای CV و VC اضافی در مسیر گفتار درج شده است. و دوم اینکه اگر فضای شناسائی همخوان یا فضای تصمیم‌گیری بتواند به تعداد کمی از همخوان‌ها به طریقی محدود بشود، صحت بازشناسی بسیار بالا خواهد رفت. و این نکته مهمی است که بنظر می‌رسد سیستم ادراک انسان نیز به کمک آن به صحت بازشناخت بالائی در واجها می‌رسد. در رابطه با این دو نکته در قسمت بررسی خطاهای مدل بیشتر صحبت خواهیم کرد.

۱-۶ـ شبکه بازشناسی نوع همخوان در موقعیت بعد از واکه (VCT)

شبکه مورد استفاده در اینجا تقریباً مشابه مورد قبلی است، با این تفاوت که این بار در ورودی ۷ فریم مربوط به همخوان در سمت راست قرار می‌گیرد و مورد دیگر نیز اینکه تعداد خروجی‌های شبکه در این حالت ۲۰ می‌باشد. علت این کاهش تعداد همخوان‌ها در اینجا این است که واج‌های انفجاری واکدار و بیواک دو بدو در قسمت ورود به واج (بست واج) یکسان هستند. به عنوان مثال ماهیت صوتی سیگنال در دو مرز واجی «ab» و «ap» مشابه است و تفاوت دو واج b و p عمدتاً در ناحیه رهش [۱۰] آن است. و لذا در خروجی این شبکه که از اطلاعات مربوط به ناحیه گذرای مرز VC نوع همخوان را مشخص می‌کند دو مرز فوق با یک خروجی مشخص می‌شوند در مواقعی که یک همخوان مابین دو واکه قرار می‌گیرد آنرا C_1 درنظر گرفته و با کمک هر دو شبکه VCT و CVT آنرا شناسائی و نتایج را باهم ترکیب می‌کنیم که در این حالت صحت بازشناسی نسبتاً بالاست. در حالاتی که دو یا سه همخوان در ما بین دو واکه قرار می‌گیرند و به عنوان C_3 و C_2 و C_1 باید شناسائی شوند ($VC_2C_3C_1V$) خطا نسبتاً زیاد است. در چنین حالتی از یک شبکه‌ای که روی مرزهای CC ثابت و منطبق شود و دو همخوان دوطرف این مرز را شناسائی نماید نیز می‌توانیم کمک بگیریم. در هر حال دراین موارد خطای بازشناسی در تعداد و محل مرزهای CC و نوع همخوان‌ها نسبتاً زیاد است.

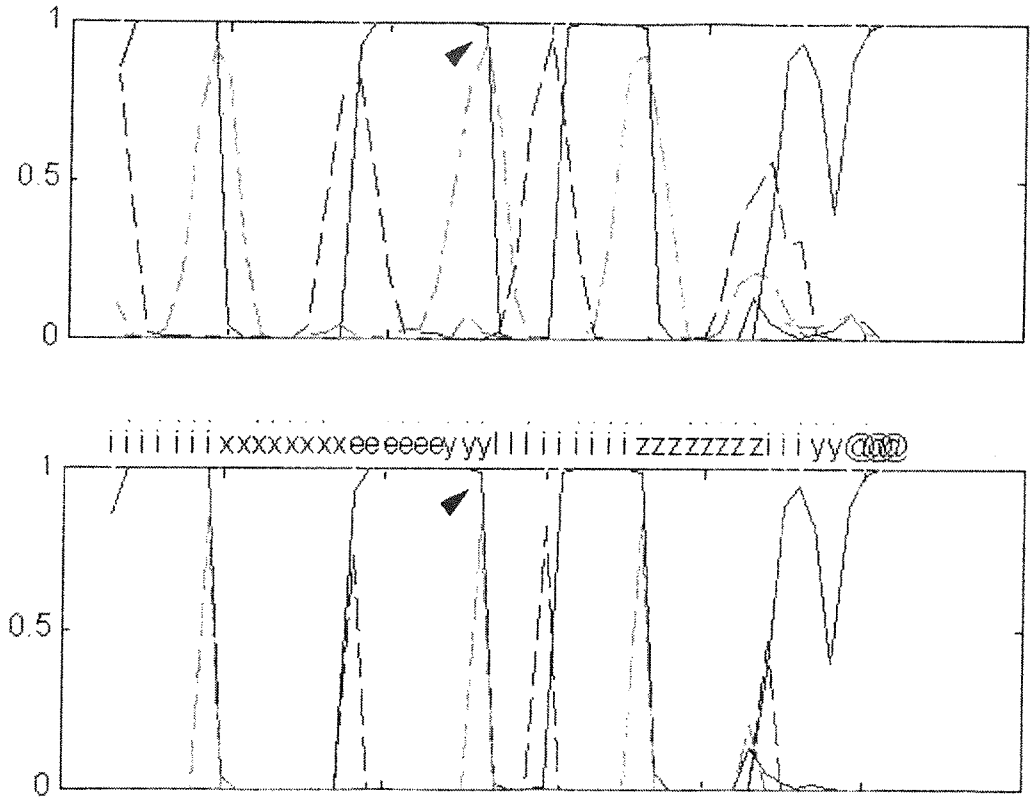
۲ـ تحلیل نتایج پیاده‌سازی مدل و آنالیز خطاهای بازشناسی

همانطور که قبلاً نیز اشاره کردیم صحت بازشناسی همخوانهائی که در مجاورت واکه قرار دارند توسط دو شبکه VCT و CVT، هنگامی که در تعیین مکان مرزهای CV و VC خطا نداشته باشیم در حدود ۸۵٪ بدست آمده است و این برای همخوان‌ها که نوعاً ضعیفتر از واکه‌ها شناسائی می‌شوند در سطح خوبی است.

مشاهده نتایج عملکرد شبکه VCS نشان می‌دهد که عملاً در مواقع مشخصی این شبکه با مشکل مواجه است. یکی از این موارد در واج‌های «i» و «y» است. دقت در محتوای صوتی این دو واج نشان می‌دهد که در نواحی ایستان سیگنال گفتار این دو واج مشابه هستند و آنچه که موجب تمیز دادن آنها از هم می‌شود به واج‌های قبل و بعد آنها برمی‌گردد. ما این نکته را هنگام اداکردن و شنیدن دو هجای «meyl» و «mil» می‌توانیم دریابیم. و لذا در نواحی ایستان، هر دو واج «y» و «i» باید در شبکه با یک خروجی بیان شود تا شبکه با تناقض روبرو نشود.

این نکته اساس مدل‌های هجائی بر مبنای C-V را تا حدی متزلزل می‌کند و نشان می‌دهند که انتساب «y» به عنوان همخوان و «i» به عنوان واکه که با توجه به ادراک انسان بوده با ماهیت کمی دو سیگنال در آنالیز مهندسی تطابق ندارد (نمودار ۳).

به عبارت دیگر این بررسی نشان می‌دهد که برخی نکات نظری زبانشناسی با نتایج پیاده‌سازی‌های کمی و مهندسی کاملاً تطبیق نمی‌کنند و نیاز است که پیاده‌سازی مهندسی نظریات زبانشناسی، به روش‌های هوشیارانه‌تری انجام پذیرد. یک نکته در این رابطه این است که عملاً مدل بازشناس بایستی قادر باشد که واج قبل و بعد را نیز درنظر بگیرد و نکته دیگر اینکه برای ترکیب اطلاعات نواحی گذرا با نواحی ایستان آواها، استفاده از مدل همخوان ـ واکه V-C چندان مناسب نیست و بهتر است که از مدل‌های بر مبنای شناسائی مرزهای مابین هر دو واج معین بصورت جداگانه استفاده کنیم که این به معنی شناسائی توأم دوآواها [۱۱] و نواحی ایستان گفتار توسط شبکه‌های عصبی خواهد بود. برای این منظور ما پیشنهاد می‌کنیم که برای عملکرد بهتر مدل، نواحی مرزی آواها باید برای هر دو واج مشخص بصورت جداگانه، ثابت و با بکارگیری روش بهترین انطباق مورد تعلیم و بازشناسی قرار گیرند و نواحی ایستان آواها نیز توسط شبکه‌ای جداگانه بصورت طی مسیر بردار بازنمائی شناسائی شوند و نتایج این‌ها برای هر آوا در یک شبکه عصبی مدولار و سلسله مراتبی بایستی ترکیب گردند. این روش این امکان را می‌دهد که اطلاعات نواحی گذرای آواها را با کیفیت بالائی حفظ و استخراج نمائیم و با اطلاعات سایر نواحی ترکیب کنیم. از سوی دیگر با محدود کردن فضای جستجو در بازشناسی مرزها به هر دو واج مجاور، امکان تعلیم و شناسائی دقیقتر در مدل بازشناسی آواها در یک فضای تصمیم‌گیری بهینه فراهم می‌گردد.



نمودار(۳)خطای شبکه VCS در طبقه بندی «i» و «y» به عنوان واکه یا همخوان.

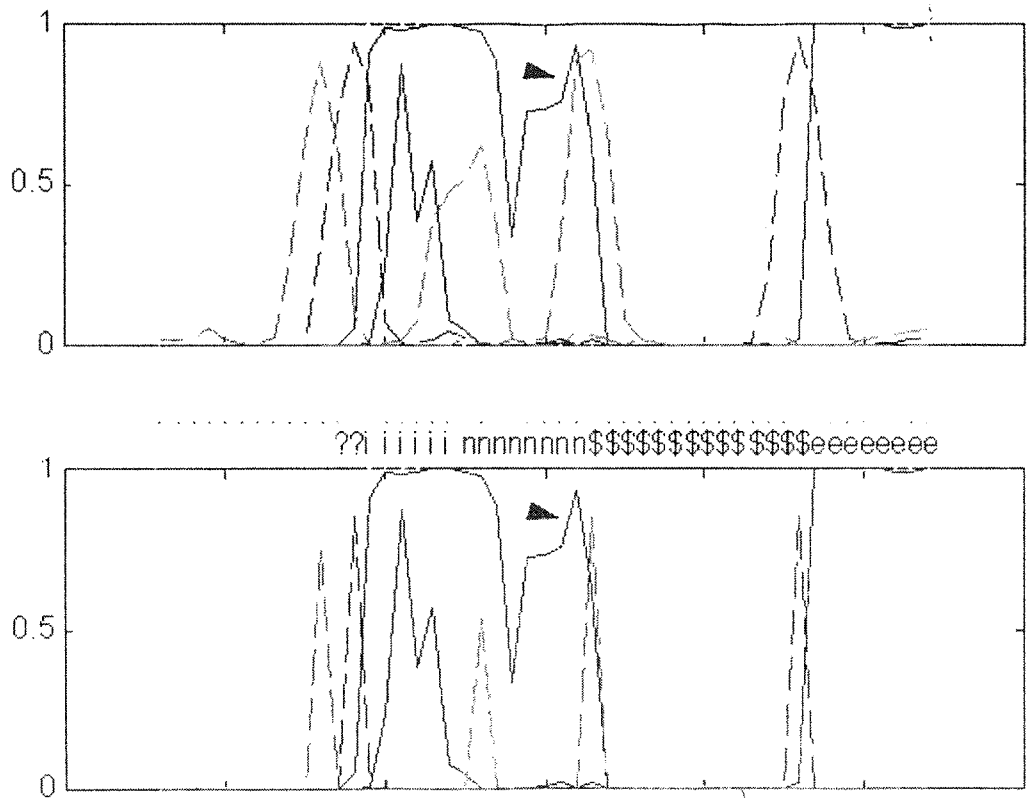
مشکل موجود در تفکیک دو واج «y» و «i» باعث می‌شود که تشخیص مرزهای CV و VC توسط شبکه VTC نیز دچار اشکال شود و مرزهای CV و VC زائد در مسیر سیگنال درج گردند. از موارد دیگر خطای موجود در عملکرد این مدل این است که گهگاه همخوان‌های شبه واکه[۱۲]، مانند «v»، «m»، «n»، «l» واکه خصوصاً «u» تشخیص داده می‌شوند و لذا مرز VC یا حذف شده و یا به بعد از این همخوانها منتقل می‌گردد. حذف مرزهای VC و یا CV در مسیر سیگنال عملکرد مدل را دچار اختلال می‌کند و موجب شناسائی غلط واجها می‌گردد. شناسائی مرز CV عمدتاً با دقت و صحت بالائی همراه است و لذا می‌تواند مبنای شناسائی هر هجا قرار گیرد. در شناسائی سایر واجها، مدل غالباً با کیفیت خوبی عمل می‌کند و واجها و هجاها را استخراج می‌نماید.در مرز ما بین همخوان‌های شبه واکه با سایر همخوانها گاهی خروجی‌های CV و یا VC فعال می‌شوند و مشابهت شبه واکه‌ها با واکه‌ها ایجاد مشکل می‌نماید. همچنین نواحی میانی شبه واکه‌های طولانی، در برخی موارد واکه شناخته شده است.

خطاهای دیگری که در عملکرد این مدل مشاهده می‌شود وجود یک واکه کوتاه بعد از همخوان‌های ساکن خصوصاً همخوان «r» و همچنین همخوان‌های «q»، «k»، «h»، «l»، «n»، «y» است (نمودار۴ و ۵).

عملاً ما در هنگام بیان این همخوان‌ها در وضعیت ساکن در کلمات غالباً یک واکه کوتاه به دنبال آن می‌آوریم. این واکه کوتاه می‌تواند موجب درج یک هجای اضافی در این مدل در مسیر سیگنال گفتار بشود. در برخی نواحی از سیگنال، همخوان «v» با واکه «u» اشتباه می‌شوند و این به دلیل نزدیک بودن وضعیت ناحیه صوتی هنگام بیان این دو واج است. نتیجه بسیار مهمی که از بررسی این خطاها می‌توان فهمید این است که سیستم ادراک گفتار انسان، آواها را فقط براساس محتوای صوتی موضعی آنها شناسائی نمی‌کند بلکه آواهای مجاور بطور ناخودآگاه در شناسائی هر آوا دخالت می‌کنند و هر آوا با توجه به موقعیت آن در زنجیره آواهای اطراف خود [۱۳] شناسائی می‌گردد. لذا در طراحی مدل‌های آتی بازشناسی آواها و هجاها به این نکته مهم بایستی توجه نمود. به همین دلیل ما برای شناسائی هر آوا، بکارگیری اطلاعات قسمتهای ایستان و گذرای اطراف آنرا در یک شبکه عصبی مدولار و سلسله مراتبی پیشنهاد کرده‌ایم.

علاوه بر این، تحلیل نتایج پیاده‌سازی ما نشان می‌دهد که اگر سیگنال صوتی را زنجیره‌ای از وقایع و حالت‌های بدنبال هم در یک پردازش بر مبنای سیستم‌های وقایع گسسته[۱۴] درنظر بگیریم این مدل با ماهیت گفتار تناسب بهتری خواهد داشت.

ضرورت تحمل تغییرات طول آوا در حین صحبت با سرعت‌های متفاوت نیز دلیلی بر تأیید این مطلب است. برای این منظور ما بکارگیری شبکه‌های عصبی با تأخیرهای زمانی متغیر و سنکرون با وقایع را پیشنهاد می‌کنیم. ما در این رابطه مدلی طراحی و پیاده‌سازی کرده‌ایم که در یک مقاله جداگانه نتایج آنرا خواهیم آورد.در چنین مدلی امکان استخراج طول زمانی آواها به عنوان



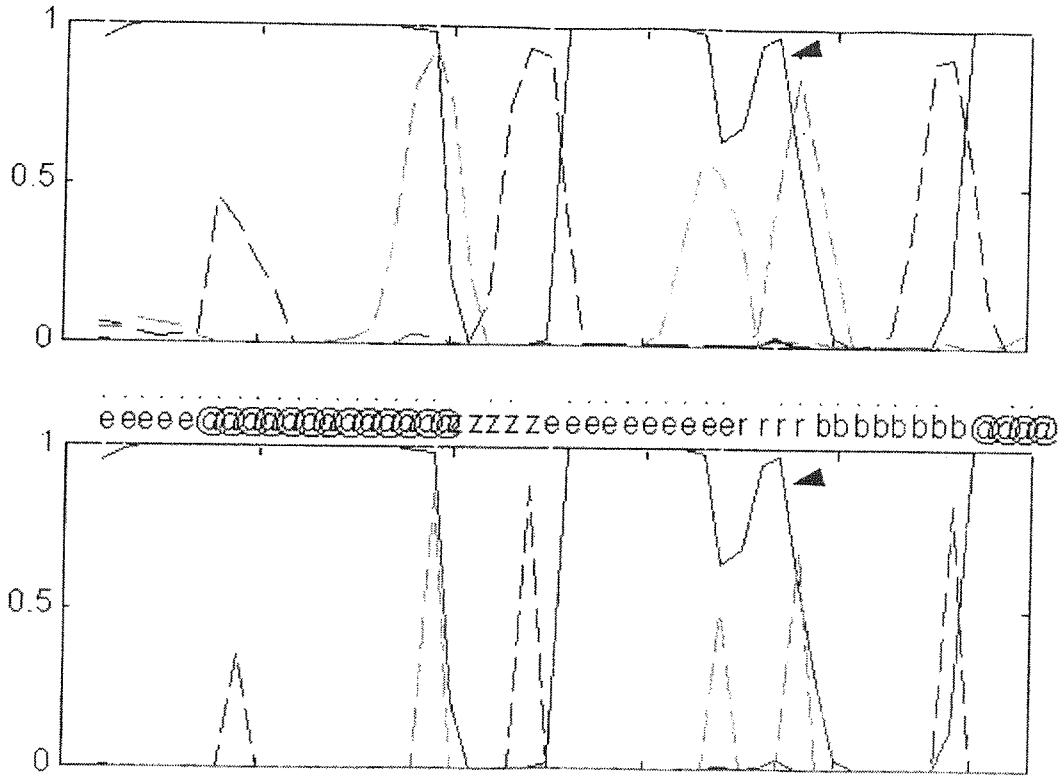
نمودار(٤) خطا در شناسایی «n» به عنوان همخوان و درج یک واکه مجازی بعد از همخوان«n» ساکن.

یک اطلاعات مفید نیز فراهم می‌شود و همچنین اطلاعات زنجیره وقایع[15] و حالات[16] گذشته می‌تواند به نحو مناسبی در جهت محدود کردن فضای جستجو و تصمیم‌گیری هنگام بازشناسی هر آوا و یا واقعه جدید بکار گرفته شود.

بوسیله روش بکارگیری شده در این مقاله همچنین می‌توان مستقیماً هجا را به عنوان واحد زبانی پایه شناسائی نمود. این عمل بوسیله وارد کردن نتایج شبکه‌های بازشناسی VCT, CVT, VS بصورت یکجا در یک ساختار شبکه عصبی واحد بصورت مدولار، امکان‌پذیر می‌گردد. نکته‌ای را هم که در اینجا می‌توان ذکر کرد این است که اگر ما بتوانیم موارد درج و حذف مرزهای CV و VC را در مدل به طریقی از بین ببریم و بدینوسیله مشکل حذف و یا افزایش هجا را حل نمائیم آنگاه با این مدل ما قادر خواهیم بود به درصد صحت بازشناسی واج و هجای بالائی دست یابیم و لذا باید به دنبال روشهائی باشیم که مانع از درج مرزهای اضافی و یا حذف مرزهای موجود در سیگنال بشوند. یک راه‌حل ممکن برای این مسئله، رفع مشکل واج‌های «y» و«i» و نیز دخالت دادن طول واکه برای تفکیک واکه‌های واقعی از واکه‌های کوتاه به غلط درج شده می‌باشد. در صورت وصول به چنین هدفی آنگاه می‌توانیم امید داشته باشیم که به صحت بازشناخت واج مستقل از گویندهٔ حدود ۸۵ تا ۹۰ درصد بتوان رسید.

۳ـ جمع‌بندی

بررسی نتایج روش‌های رایج بازشناسی گفتار نشان داده‌اند که اطلاعات گفتاری مفید و قابل ملاحظه‌ای را از نواحی گذرای مرزی آواها می‌توان استخراج کرد. برای این منظور ما از شبکه‌های عصبی بصورت ثابت شده بر روی مرز و با بکارگیری روش بهترین انطباق استفاده می‌کنیم تا تلفات اطلاعات را در مرزها به حداقل برسانیم. سپس نتایج بازشناسی نواحی گذرای مرزی



نمودار(۵) درج واکه مجازی بعد از همخوان «r» ساکن.

آواها را با نتایج شناسائی نواحی ایستان آواها به نحو مناسب ترکیب می‌کنیم. برای مشخص کردن مرزهای واکه‌ها و همخوان‌ها ما از شبکه عصبی با خروجی ضربه‌ای استفاده کرده‌ایم. ترکیب مناسب نتایج ما را به صحت بازشناسی واج برتری می‌رساند. برای عملکرد بهتر این مدل ما درنظر داریم که بجای تشخیص مرزهای واکه‌ها و همخوان‌ها، مرزهای مربوط به هر دو واج معین را شناسائی کنیم و به این وسیله خطای عملکرد مدل را کاهش دهیم. ترکیب مناسب نتایج بازشناسی وقایع صوتی متوالی می‌تواند با محدود کردن فضای جستجو و تصمیم‌گیری، صحت بازشناسی بالاتری را به ما بدهد. بکارگیری روش‌های مناسب جلوگیری از درج و حذف واکه‌ها و نیز مرزهای CV و VC زائد، می‌تواند به ما کمک کند تا به درصد صحت بازشناسی واج مستقل از گویندهٔ بالائی در حدود ۸۵٪ برسیم.

۵ـ قدردانی

این تحقیق با حمایت و همکاری پژوهشکدهٔ پردازش هوشمند علائم به انجام رسیده است.

زیرنویسها

۹ ـ دراین نامگذاری بلوکها با حروف مخفف V، برای واکه،C برای همخوان، S برای نواحی یکنواخت وT برای نواحی گذرا مورد استفاده قرار گرفته‌اند.

1-Phoneme
2-Segmentation
3-Maximum
4-Minimum
5-Articulators
6-Syllabic
7-Impulsive
8-Bark Scale
10-Release
11-Diphones
12-Semi-Vowels
13-Context
14-Discrete event systems
15-Events
16-States

مراجع

[1] Glaeser, A. "Modular Neural Networks with Task-Specific Input Parameters for Speaker Independent Speech Recognition", Proceedings of the ESCA. EUROSPEECH'95, pp. 1655 - 1658, Madrid, (1995).

[۲] سید صالحی، س.ع، بازشناخت گفتار پیوسته فارسی با استفاده از مدل عملکردی مغز انسان دردرک گفتار، پایان‌نامه دکتری، دانشگاه تربیت مدرس، تهران، ۱۳۷۴.

[۳] سید صالحی، س.ع. و همکاران، بررسی عملکرد شبکه‌های عصبی TDBP در بازشناخت مستقل از گوینده واج‌های فارسی، مجموعه مقالات کنفرانس بین‌المللی سیستم‌های هوشمند و شناختی، ص ۲۴-۲۹، تهران، ۱۳۷۵.

[۴] سید صالحی، س.ع. و همکاران، بازشناسی مستقل از گوینده واج‌های فارسی با استفاده از مشخصه‌های تولیدی، مجموعه مقالات کنفرانس بین‌المللی سیستم‌های هوشمند و شناختی، ص ۹۳-۹۷، تهران، ۱۳۷۵.

[5] Yu, H.-J., oh, Y.-H. "A Neural Network Using Nonuniform Units for Continuous Speech Recognition", Proc. of EUROSPEECH'95, Vol. 3, pp. 1677-1680, Madrid (1995).

[6] Yu, H.-J., oh, Y.-H. "A Neural Network Using Acoustic Sub-word Units for Continuous Speech Recognition" , Proc. of ICSLP'96, pp. 506-509 (1996).

[7] Morgan, N., et al. "Stochastic Perceptual Auditory-Event-Based Models for Speech Recognition", Proc. of ICSLP'94, pp. 1943-1946, Yokohama (1994).

[8] Duez, D. "Effects of Articulation Rate on Duration in Read French Speech", Proc. of EUROSPEECH'99-CD, Budapest (1999).

[۹] بی‌جن‌خان، م.، بازنمائی واجی و آوائی زبان فارسی و بکارگیری آن درباز شناخت خودکار گفتار، پایان‌نامه دکتری،گروه زبان‌شناسی، دانشگاه تهران، تهران، ۱۳۷۴.

[۱۰] سید صالحی، س.ع. و همکاران، ارزیابی عملکرد شبکه‌های عصبی در بازشناسی مستقل از گویندهٔ واج‌ها با استفاده از چند بازنمائی الهام گرفته از سیستم شنوائی، مجموعه مقالات سومین کنفرانس الکترونیک، ص ۱۱۴-۱۲۵، دانشگاه شیراز، ۱۳۷۴.

[11] Bijankhan, M., et al. "FARSDAT-The Speech Database of Farsi Spoken Language" , Proceedings of SST-94, pp. 826-831, Perth, (1994).