بازشناسی اعداد فارسی بر روی خط تلفن:
مقایسه‌ای بین روشهای آماری، عصبی و روش هیبرید

محمدمهدی همایون‌پور
استادیار

جهانشاه کبودیان
دانشجوی دکتری

حسین بشیری
کارشناس ارشد

ذبیح‌اله احمدپور
کارشناس ارشد

دانشکدهٔ مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی امیرکبیر

چکیده

این مقاله به ارائه نتایجِ تحقیقاتِ صورت گرفته برای بازشناسی ارقام و اعداد فارسی که بصورت گسسته، متصل و پیوسته از طریق تلفن بیان شده‌اند اختصاص دارد. در این مقاله تأثیر انتخاب واحد صوتی برای بازشناسی (واج یا کلمه) بررسی شده و مقایسه‌ای بین روشهای دسته‌بندی آماری مانند مدل مخفی مارکوف، شبکه عصبی چند لایه پرسپترون و هیبرید صورت گرفته است. آزمایشها نشان می‌دهند که در محیط نویزی و تلفنی، کارایی پارامترهای کپسترال حاصل از بانک فیلتر بیش از کارایی پارامترهای کپسترال حاصل از آنالیز پیشگویی خطی است. همچنین مشاهده گردید که کارایی مدل پنهان مارکف پیوسته هم از شبکهٔ عصبی و هم از مدل هیبرید بیشتر می‌باشد. در نهایت، بهترین راندمان برای بازشناسی ارقام گسسته و ارقام متصل(با مدلهای کلمه‌ای)، بر روی پایگاه دادهٔ تلفنی FARSDIGITS1 و بازای داده‌های آزمایشی بترتیب برابر با ۹۹/۱٪ و ۸۳/۷٪ بوده و بهترین راندمانِ شناساییِ کلمات در بازشناسی اعداد پیوسته بصورت نامقید (با مدلهای واجی)، بر روی پایگاه دادهٔ تلفنی FARSNUMBERS1 و بازای داده‌های آزمایشی برابر با ۹۱/۱٪ بوده است.

کلمات کلیدی

بازشناسی گفتار، بازشناسی ارقام گسسته، بازشناسی ارقام متصل، بازشناسی اعداد پیوسته، مدل پنهان مارکُف، شبکه عصبی، تفاضل طیفی، روش بازتخمین نهفته، مدل هیبرید.

Recognition of Farsi Number over Telephone: A Comparison of Statistical, Neural, and Hybrid Approaches

M. M. Homayounpour
Assistant Professor

J. Kabudian
Ph.D. Student

H. Bashiri
MSc. Graduate

Z. Ahmadpour
MSc. Graduate

Department of Computer Engineering and Information Technology,
Amirkabir University of Technology

Abstract

This paper presents results of research on isolated, connected and continuous farsi number recognition over telephone. Effect of selecting basic recognition unit (phoneme or word) is investigated, and statistical, neural and hybrid approaches are compared. In farsi digit and number recognition, continuous density hidden markov models with word and phoneme models, embedded reestimation after baum-welch, NN and hybrid methods were used. Experimental results show that in noisy and telephony conditions, the system employing cepstral vectors derived from mel-frequency filter-bank analysis outperforms the system employing cepstral vectors derived from linear-prediction analysis. also, it was observed that performance of HMM system is more than that of both the NN and hybrid systems. The best performance of digit recognition in isolated and connected farsi digit recognition (with word models) over farsdigits1 telephony database was %99.1 and %83.7 respectively, and the best performance of word recognition in unconstrained farsi number recognition (with phoneme models) over farsnumbers1 telephony database was %91.1.

Keywords

Speech Recognition, Number Recognition, Isolated Digit Recognition, Connected Digit Recognition, Hidden Markov Model, Multi-Layer perceptron, Hybrid MLP/HMM System

مقدمه

همگانی شدن کامپیوتر و امکان دسترسی به آن چه بطور مستقیم و چه از راه دور از طریق تلفن و اینترنت موجب شده است که بسیاری از خدمات شامل دسترسی به اطلاعات، بانک‌های داده، سرویس‌های بانکی و نیز بهره‌برداری از بسیاری از خدمات دیگر بتواند از طریق کامپیوتر صورت گیرد. صوت و بطور خاص گفتار یکی از راحت‌ترین و قابل دسترس‌ترین راههای برقراری ارتباط انسان با کامپیوتر جهت دسترسی به اطلاعات و خدمات می‌باشد. بازشناسیِ خودکارِ گفتار نوعی فناوری است که به یک کامپیوتر این امکان را میدهد که گفتار و کلماتِ گوینده‌ای را که از طریق میکروفون یا پشت گوشیِ تلفن صحبت می‌نماید، بازشناسی کند. بازشناسی اعداد توسط کامپیوتر که بصورت گفتاری بیان شده باشند، بدلیل مورد استفاده بودن در بسیاری از کاربردها از جمله اعلام شمارهٔ شناسایی شخصی، کد شناسایی ملی، شمارهٔ حساب بانکی یا شمارهٔ عضویت برای کاربران یک سیستمِ خدمات‌رسانی، ثبت‌نام دانشجویان از طریق تلفن یا اینترنت پس از اعلام شمارهٔ دانشجویی یا کد دروس مورد نظر و مانند آن بسیار حائز اهمیت می‌باشد.

یک سیستم بازشناسیِ ایده‌ال بایستی بتواند ۱۰۰٪ گفتارِ هر گوینده‌ای را که با وضوحِ کافی صحبت می‌نماید، مستقل از حجم مجموعهٔ لغات، نویز، لهجه و مشخصات گویشیِ گوینده و نیز خصوصیاتِ کانالهای مخابراتی که گفتار از طریق آنها منتقل شده است، تشخیص دهد. البته بلادرنگ بودن در بازشناسی گفتار از اهمیت بالایی برخوردار می‌باشد. با وجود آنکه چند دهه از تحقیقات در زمینهٔ پردازش گفتار می‌گذرد، با این وجود راندمانِ بالای ۹۵٪ در بازشناسی گفتار تنها در مواقعی قابل دسترسی است که محدودیتهای ویژه‌ای را در بازشناسی و شرایط آن لحاظ نماییم. با توجه به چگونگی و میزان این محدودیتها، درجات مختلفی از راندمان در بازشناسی گفتار قابل دسترسی است. برای بازشناسی گفتار پیوسته با مجموعهٔ لغات بزرگ و با فرضِ انتقال گفتار از طریق کانالهای مخابراتی، برای رسیدن به راندمان مطلوب همچنان به تحقیقات بیشتری نیاز می‌باشد. این در حالی است که زمانِ بازشناسی در این سیستمها در حال حاضر از زمان بلادرنگ بیشتر می‌باشد. لازم بذکر است که بازشناسی گفتار چنانچه از راه دور و از طریق خط تلفن بیان شده باشد، بدلیل تأثیرات ناشی از گوشی تلفن و نیز خطوط مخابراتی، از پیچیدگی و ویژگیهای خاص خود برخوردار می‌باشد.

تکنیکِ غالبِ مورد استفاده در بازشناسیِ اتوماتیک گفتار، روش مدل مخفی مارکف HMM می‌باشد. بازشناسی گفتار میتواند مبتنی بر کلمه، واج یا مانند آن باشد. HMM با بیشینه کردن میزان درستنماییِ (Likelihood) هر واج یا کلمه، گفتار را بازشناسی می‌نماید. چنانچه واحد بازشناسیِ گفتار، واج باشد، هر کلمه‌ای در لیستِ مجموعهٔ لغات بر اساس واجهای تشکیل‌دهندهٔ آن تعریف می‌شود. یک روندِ جستجو بنام الگوریتم ویتربی برای تعیینِ دنبالهٔ واجها با بیشترین میزانِ درستنمایی

بکار گرفته میشود. جستجو تنها به یافتن دنباله‌هایی از واجها که متناظر با کلماتِ موجود در لیستِ مجموعهٔ لغات هستند، منحصر می‌باشد و دنباله‌ای از واجهای با بیشترین درستنماییِ کلی مشخص کنندهٔ کلمهٔ گفته شده می‌باشد. در HMM استاندارد مقادیر درستنمایی با استفاده از تابع چگالی احتمالِ مخلوط گوسی محاسبه میشود. در بعضی از مواردیکه HMM را بصورت ترکیب با شبکه‌های عصبی بکار میبریم، تابع چگالی احتمال به کمک شبکه‌های عصبی تخمین زده میشود.

بازشناسی گفتار معمولاً به سه دستهٔ بازشناسی کلمات گسسته، کلمات متصل و نیز گفتار پیوسته تقسیم‌بندی میشود. در بازشناسی کلمات گسسته فرض بر آن است که گوینده کلمات را بصورت جدا ـ جدا و با مکثِ حداقل ۲۰۰ میلی ثانیه بین آنها بیان می‌نماید. در بازشناسی کلمات متصل، مکثِ بین کلمات ضروری نیست، لیکن از گوینده خواسته می‌شود که کلمات را بصورت مؤکد و واضح بیان نماید. در بازشناسیِ گفتار پیوسته هیچ نوع شرط خاصی در بیان گفتار و کلمات لحاظ نمیگردد. بازشناسیِ کلمات متصل نسبت به بازشناسی کلمات گسسته، کاری مشکل‌تر و پیچیده‌تر می‌باشد. اثرگذاریِ کلمات بر روی همدیگر هنگامِ آدا شدن توسط گوینده‌ها و حتی حذف‌شدنِ واجهای ابتدایی و انتهایی هنگام چسبیدن کلمات به هم کار بازشناسی را مشکل میسازد. مشکلات موجود در بازشناسی کلمات متصل در بازشناسی گفتار پیوسته نیز مطرح بوده بلکه شدیدتر می‌باشد، چرا که در بازشناسی گفتار پیوسته تأثیر متقابل آواها بر یکدیگر به مراتب بیشتر بوده و چون گوینده ملزم به بیان مؤکد و واضح کلمات نمی‌باشد، امکان بیان ضعیف و حتی حذف بعضی از آواها بسیار زیاد می‌باشد.

چنانچه محدوده کلمات مورد بازشناسی را به ارقام و اعداد محدود کنیم، بازشناسی را بازشناسی اعداد می‌نامیم. در اینجا منظور از عدد مجموعه‌ای از یک یا چند رقم است که بصورت دلخواه بیان شوند. در این کار تحقیقاتی کاربر می‌تواند بعنوان مثال عدد ۱۳۵۴۷ را بصورت "یک، سه، پنج، چهار، هفت" ، بصورت "سیزده، پنجاه و چهار، هفت"، بصورت "سیزده، پانصد و چهل و هفت"، بصورت "صد و سی و پنج، چهل و هفت"، بصورت "سیزده هزار و پانصد و چهل و هفت" و یا بصورتهای دیگر بیان نماید. بازشناسی اعداد بدلیل استفاده از آن در بسیاری از زمینه‌ها و کاربردها از اهمیت بالایی در سیستمهای بازشناسی گفتار برخوردار می‌باشد. بعنوان چند نمونه از کاربردهای بازشناسی اعداد، میتوان به اعلام شمارهٔ شناسایی شخصی، کد شناسایی ملی، شمارهٔ حساب بانکی یا شمارهٔ عضویت برای کاربرانِ یک سیستم اطلاع رسانی برای دسترسی به پایگاههای داده، خرید و تجارت الکترونیکی و ثبت‌نام دانشجویان پس از اعلام شمارهٔ دانشجویی یا کُد دروس مورد نظر از طریق میکروفون، تلفن یا اینترنت اشاره کرد. در بازشناسی اعداد لازم است که گوینده نسبت به بیان دنباله‌ای از ارقام و اعداد اقدام نماید. بعنوان مثال در هنگام انجام ثبت‌نام دانشجویی لازم است که دانشجو شمارهٔ دانشجویی خود را اعلام نماید. این شمارهٔ دانشجویی میتواند بصورت ارقام گسسته، متصل یا پیوسته بیان گردد. بدلیل اهمیت بازشناسی اعداد در کاربردهای مختلف که به چند نمونه از آنها اشاره شد، این تحقیق به بازشناسیِ اعداد فارسی که بصورتهای گسسته، متصل و پیوسته بیان شده باشند، اختصاص دارد.

در زمینهٔ بازشناسی ارقام و اعداد فارسی، کارهای تحقیقاتی متعددی انجام گرفته است که از آن جمله می‌توان به موارد زیر اشاره کرد. در سال ۱۳۷۱ فرامرز فکری و همکاران [۱] با استفاده از مدل پنهان مارکف گسسته یک سیستم بازشناسی ارقام مجزای فارسی را در محیط بدون نویز و میکروفنی پیاده‌سازی نمودند و راندمان ۹۶/۰۹٪ را برای داده‌های آزمایشی بدست آوردند. در سال ۱۳۷۸ حسن بابابیک [۲] با استفاده از تلفیق مدل پنهان مارکف و شبکه عصبی برای بازشناسی ارقام مجزای میکروفنی راندمان ۹۸/۵٪ را بدست آورد. در تحقیق دیگری که در سال ۱۳۷۸ توسط آقایان سعید بابایی‌زاده و همکاران [۳] انجام گرفت، مدل ترکیبی شبکه‌های عصبی و مدل پنهان مارکف گسسته پیاده‌سازی شد. این سیستم بر روی ارقام صفر تا نه، کلمات بله و خیر که بصورت میکروفنی بیان شده بودند بکار رفت. ۱۴ مدل HMM تولید و خروجی این مدل‌ها برای کم کردن خطای تصمیم‌گیری به شبکهٔ MLP داده شد. این سیستم بر روی پایگاهی از ۲۳۰ گوینده آزمایش گردید که در حالت HMM گسسته، نرخ بازشناسی آن ٪۹۸ و برای حالت ترکیبی این نرخ ۹۷/۹٪ بوده است. در سال ۱۳۷۷ شیوا رستم زاده و همکـــاران [۴] با استفاده از مدل پنهان مارکف پیوسته با ۶ حالت بازاء هر کلمه و با ۲۰۰ نمونه برای آموزش و ۴۰ نمونه برای تست در محیط کنترل شده و میکروفنی به راندمان ۹۹/۷۵٪ برای بازشناسی ارقام گسسته دست یافتند. ضرایب مورد استفاده، ضرایب LPCC با ابعاد ۱۴ بوده است. در سال ۱۳۷۸ آقای امیر نجاری و همکاران[۵] با استفاده از مدل پیشگوی عصبی و با ۴۰ گوینده برای آموزش و ۱۰ گوینده برای تست در محیط تلفنی به راندمان ۸۱٪ رسیده‌اند. آموزش این سیستم با استفاده از ترکیبی از الگوریتم برنامه‌ریزی پویا و آموزش شبکه عصبی انجام می‌گرفت. این سیستم بر روی پایگاه دادهٔ تلفنی متشکل از ارقام صفر تا

نه آزمایش گردید و از ۱۶ ضریب MFCC استفاده شد. همچنین در سال ۱۳۷۹، تحقیقی توسط آقایان ابوالقاسم صیادیان و همکاران [۶] انجام شد که در آن از مدل HMM تک حالته یا بعبارتی از مدل GMM استفاده گردید. در این سیستم تعداد مخلوط گاوسی ۸ در نظر گرفته شد. این سیستم با یک سیستم HMM پیوسته با ۵ حالت در هر مدل و ۱۶ مخلوط گاوسی مقایسه گردید. با در نظر گرفتن ضرایب MFCC و مشتق آن و مشتق انرژی، نرخ بازشناسی برای سیستم پیاده‌سازی شده ۱۰۰٪ و برای HMM پیوسته ۹۴/۱۷٪ بوده است. در سال ۱۳۷۹ سیستمی توسط آقایان احمد اکبری و همکاران [۷] پیاده‌سازی گردید که در آن مقایسه‌ای بین سیستم HMM گسسته و HMM پیوسته صورت گرفت. در این تحقیق برای هر مدل ۶ حالت در نظر گرفته شد و بر روی پایگاهی متشکل از اعداد دو رقمی فارسی آزمایش گردید. در این سیستم هجا به عنوان واحد پایه در نظر گرفته شد. در حالت استفاده از ضرایب MFCC و انرژی و مشتق این ضرایب، راندمان برای حالت پیوسته ۹۸/۷٪ و برای حالت گسسته ۸۹/۷٪ بوده است و در حالت استفاده از ضرایب LPCC و انرژی و مشتق آنها، این نرخ به ترتیب به ۹۲/۱٪ و ۸۰٪ تقلیل یافته است. در سال ۱۳۸۰ آقای علی طاهری دمنه و همکاران [۸] اقدام به بازشناسی گفتار پیوسته فارسی توسط سیستم هیبرید متشکل از HMM و شبکه عصبی نموده و به راندمان ۷۵٪ در بازشناسی کلمه دست یافتند. در زمینه بازشناسی ارقام متصل فارسی بطور مستقل از گوینده می‌توان به کار آقای فرامرز فکری [۱] اشاره نمود که در محیط کنترل شده و با صداهای میکروفنی انجام شده است، متأسفانه هیچ نتیجه‌ای از راندمان سیستم پیاده‌سازی شده گزارش نشده است.

در تحقیقات فوق، بازشناسی اعداد عموماً بر روی ارقام گسسته و یا اعداد دو رقمی صورت گرفته است و بیشتر از دادگان‌های غیر تلفنی استفاده شده است. بازشناسی اعداد پیوسته تلفنی که خواندن اعداد بتواند به صورتهای گوناگونی که برای آن متصور است انجام شود، در بین تحقیقات صورت گرفته برای بازشناسی گفتار فارسی مشاهده نگردید. این تحقیق به بازشناسی اعداد فارسی که بصورت ارقام گسسته، ارقام متصل و اعداد پیوسته‌ای که به هر صورتی بتوانند خوانده شوند برای گفتار تلفنی و بصورت مستقل از گوینده اختصاص دارد. در این تحقیق مدل مخفی مارکف و شبکه‌های عصبی به منظور انجام بازشناسی اعداد فارسی بکار گرفته شده‌اند و ضمن مقایسه با یکدیگر، سیستم هیبرید متشکل از هر دو آنها نیز طراحی، پیاده‌سازی و ارزیابی شده است.

ادامه مطالب این مقاله به شرح زیر می‌باشد. در بخش ۲ بطور مختصر به خصوصیات خط تلفن اشاره خواهیم کرد. در بخش ۳ الگوریتم استفاده شده برای تشخیص گفتار از سکوت تشریح می‌شود. بخش ۴ به حذف نویز از سیگنال بروش تفاضل طیفی اختصاص دارد. در بخش ۵ استخراج ویژگی و در بخش ۶ روش تفاضل میانگین در حوزه کپسترال به منظور حذف اثر کانال انتقال مورد بررسی قرار می‌گیرند. در بخش ۷ اشاره مختصری به مدل مخفی مارکف می‌نمائیم. بازشناسی ارقام گسسته در بخش ۸، بازشناسی ارقام متصل در بخش ۹ و بازشناسی اعداد پیوسته در بخش ۱۰ ارائه میگردند. بخش ۱۱ به نتیجه‌گیری و جمع‌بندی مطالب این مقاله می‌پردازد.

۱ـ خصوصیات خط تلفن و مکالمات تلفنی

پردازش سیگنالهای صوتی و گفتاری عبور داده شده از خط تلفن و نیز مکالمات تلفنی بسیار متفاوت از پردازش در محیط‌های میکروفنی و بدون نویز است. پهنای باند خطوط تلفن محدود می‌باشد و بعنوان مثال محدودهٔ ۲۰۰ Hz تا ۳۴۰۰ Hz درنظر گرفته می‌شود که بسیاری از اطلاعات مفید سیگنال گفتار را از بین می‌برد. این پدیده در سیستم‌های بازشناسی گوینده که اطلاعات گوینده در فرکانسهای بالا از اهمیت خاصی برای تمایز گویندگان برخوردار است، بیشتر اثر خود را نشان می‌دهد. بر روی خط تلفن پژواک وجود دارد. مشخصهٔ کانال تلفنی در باند عبور، یک مشخصهٔ هموار نیست و در فرکانسهای مختلف میزان تضعیف یا تقویت آن متفاوت است که این امر نیز کار بازشناسی را مشکل تر خواهد کرد. نکتهٔ بسیار مهمی که در مورد مکالمات تلفنی وجود دارد این است که گویندگان مختلف از دهنی‌های متفاوت در دستگاه تلفن خود استفاده می‌نمایند و پاسخ فرکانسی دهنی‌های مختلف ممکن است بسیار ناهموار و بسیار متفاوت از همدیگر باشد. در مرجع [8] نشان داده شده است که تضعیف یا تقویت در مشخصهٔ فرکانسی یک دهنی، در باند تلفنی ممکن است تا 25 dB تغییر داشته باشد و

بعنوان مثال دو دهنی یکی از نوع Electret و دیگری از نوع کربنی با هم مقایسه شده‌اند که مشخصهٔ فرکانسی آنها بسیار متفاوت از همدیگر و بسیار ناهموار است. علاوه بر مسائل فوق، اگر یک گوینده فقط از یک دهنی هم استفاده کند، در زمانهای متفاوت هیچ تضمینی وجود ندارد که مشخصهٔ کانال در جلسات مختلف و در تماسهای مختلف یکسان باشد. بر روی خط تلفن نویز نیز وجود دارد که لزوماً نویز جمع شونده نیست. در مکالمات تلفنی وضعیت قرار گرفتن دهان گوینده نسبت به دهنی، در مقایسه با مکالمات میکروفنی کنترل شده، تغییرات بیشتری دارد. در بعضی از دهنی‌ها، مثل دهنی کربنی، اعوجاجات هارمونیکی ایجاد میگردد و حتی پاسخ فرکانسی دهنی متغیر با زمان می‌باشد. نویزهای دیگری نیز بر روی خط تلفن وجود دارد که از آن جمله می توان به نویز آکوستیکی زمینه، نویز برق شهر، نویز همشنوایی، نویز حاصل از ارتباطات مایکروویو و ... اشاره کرد. ملاحظه میشود که مسأله شناسایی گفتار با مکالمات تلفنی بسیار متفاوت‌تر و دشوارتر از شناسایی گفتار در محیط بدون نویز، میکروفنی و فقط با یک میکروفن است. به علت وجود پدیده های فوق، باید برای سیستم‌های بازشناسی بر روی خط تلفن تمهیداتی بیندیشیم که در این مقاله برای کاهش اثرِ نویزهای جمع شونده و نیز برای جبران‌سازی مشخصهٔ کانالِ تلفنی راه حلی درنظر گرفته شده و نتایج آن در بخشهای بعد ارائه گردیده است.

۲ـ تشخیص گفتار از سکوت

برای تشخیص گفتار از سکوت زمینه از الگوریتم ذکر شده در مرجع [9] استفاده گردید. تشخیص گفتار از سکوت به منظور تعیین محدوده ارقام گسسته و نیز تشخیص نویز زمینه از بخش سکوت به منظور استفاده در بخش حذف نویز کاربرد دارد. سیگنال صحبت نمونه برداری شده با فرکانس ۱۱۰۲۵ هرتز به فریم هایی که طول هر فریم ۱۰۲۴ نمونه است و با هم همپوشانی دارند تقسیم می‌شود. هر قاب از سیگنال در یک پنجره همینگ ضرب می‌شود و تبدیل فوریه آن بدست می‌آید. برای حذف سه هارمونیک اول برق شهر و نیز حذف فرکانسهای بالای ناکارآمد در این قسمت، یک فیلتر میان گذر با باند عبور ۱۹۵ تا ۳۸۵۰ هرتز به سیگنال اعمال می‌گردد. در این مرحله اگر طیف سیگنال را X_i بنامیم، این طیف را بصورت صعودی مرتب می‌کنیم و طیف مرتب شده را Z_i می‌نامیم. مبنای تصمیم‌گیری در این الگوریتم، یک تخمین مقاوم از SNR است که برای بدست آوردن آن بصورت زیر عمل می‌کنیم. اگر N_P تخمینی از چگالی توان نویز و S_P تخمینی از چگالی توان گفتار باشد آنگاه N_P و S_P را به این شکل تعریف می‌کنیم:

$$N_P = \frac{1}{100}\sum_{i=45}^{145} Z_i \qquad (۱)$$

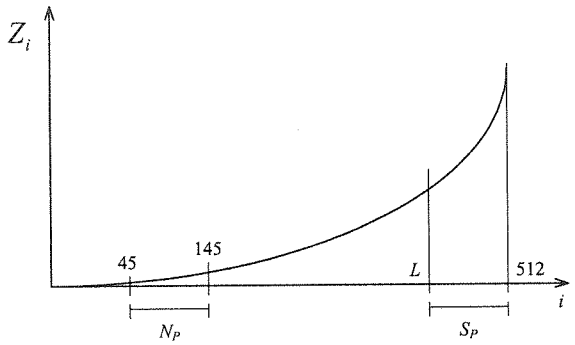
$$S_P = \frac{1}{512-L}\sum_{i=L}^{512} Z_i \qquad (۲)$$

که در آن Z_i، i امین عنصر از طیف مرتب شده بوده و L باید در رابطه زیر صدق کند:

$$E_T = \sum_{i=1}^{512} Z_i \ , \ \sum_{i=L}^{512} Z_i \geq 0.4E_T \qquad (۳)$$

که E_T انرژی کل طیف می‌باشد. شکل زیر نمایی از N_P و S_P را نشان می‌دهد. پس از این SNR را از رابطه زیر محاسبه می‌نمائیم:

$$\mathrm{SNR} = \frac{S_P}{N_P} \qquad (۴)$$

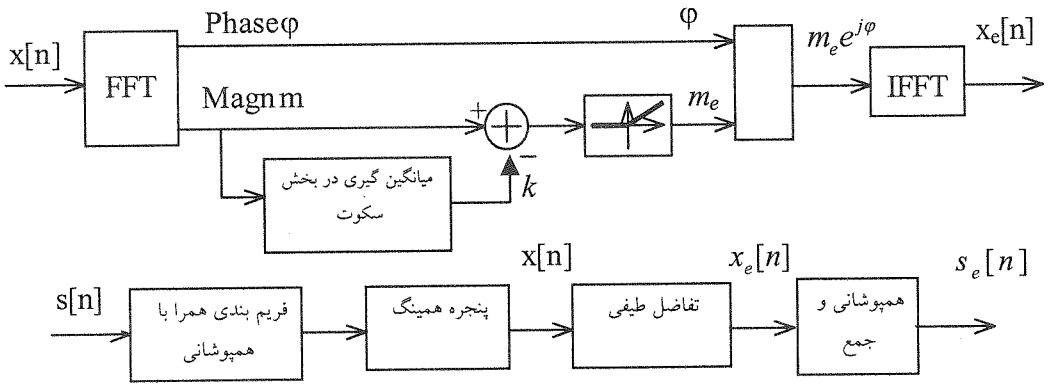


شکل (۱) تعریف S_P و N_P

چنین فرمولی برای SNR، از گین سیگنال و نیز از اندازهٔ فریم مستقل می‌باشد. SNR بدست آمده را با یک مقدار سطح آستانه مقایسه می‌کنیم. اگر SNR از مقدار سطح آستانه بیشتر بود، گفتار است و در غیر اینصورت سکوت زمینه می‌باشد. مقدار سطح آستانه برابر ۹۰ در نظرگرفته می‌شود. ادعا شده است که روش فوق حتی در SNR= 0dB هم به خوبی کار می‌کند [۹].

۳ـ حذف نویز از سیگنال صحبت

برای حذف نویز از روش تفاضل طیفی استفاده شده است[10]. بلوک دیاگرام روش تفاضل طیفی در شکل زیر نشان داده شده است.



شکل (۲) بلوک دیاگرام روش تفاضل طیفی.

این روش از روشهای بهبود گفتار با فرضِ جمعی بودنِ نویزِ تأثیرگذار می‌باشد. با فرض آنکه نویز دارای میانگین صفر بوده و با سیگنال غیر همبسته باشد، طبق شکل و رابطه زیر طیف توان نویز در بخش سکوت بدست آمده و از طیف توان سیگنال نویزی شده کم میگردد. برای تشخیص محدوده‌های سکوت از الگوریتم مذکور در قسمت قبل استفاده شده است. همانطور که مشاهده می‌شود حاصل این تفاضل ممکن است منفی شود، بعضی روشها در این شرایط مقدار منفی را صفر می‌کنند و بعضی روشهای دیگر نیز ممکن است در صورت منفی شدنِ حاصل، علامت آنرا تغییر دهند و یا مقدار آنرا برابر یک مقدار کوچک مثبت قرار دهند. در حالت کلی و بطور تقریبی می‌توان این معادله را بصورت زیر نوشت:

$$\left|S_e(w)\right|^2 = \left|Y(w)\right|^2 - k.E\left[\left|D(w)^2\right|\right] \qquad (۵)$$

در رابطه فوق Y(w)، D(w) و S(w) بترتیب طیف سیگنال نویزی، طیف نویز و طیف سیگنال بدون نویز می‌باشند. پس با استفاده از این معادلات تخمینی از مقدار $|S(w)|$ بدست می‌آید. برای پیدا کردن $\angle S(w)$ یعنی فازِ سیگنال بدون نویز، می‌توان از این ایده بهره گرفت که گوش انسان به فاز سیگنال حساس نیست. از این رو $\angle S(w)$ را با استفاده از $\angle Y(w)$

یعنی فاز سیگنال نویزی بدست می‌آورند. بنابراین تقریبی از طیف فوریهٔ سیگنال بدون نویز از رابطهٔ زیر بدست می‌آید:

$$S_e(w) = |S_e(w)|.\exp(j\angle Y(w)) \qquad (۶)$$

و با استفاده از تبدیل فوریهٔ معکوس، $s_e(n)$ بدست خواهد آمد. برای بهسازی سیگنالِ گفتار نویزی، آنرا به فریم‌هایی که با هم همپوشانی دارند، قطعه‌بندی می‌کنیم، بر روی هر فریم پنجرهٔ هنینگ را اعمال کرده، روشِ تفاضل طیفی را بر روی فریم انجام داده و سپس سیگنال زمانیِ حاصل از فریم‌های مختلف را با روش همپوشانی و جمع به هم پیوند می‌زنیم.

٤ـ استخراج ویژگی

ویژگیهای استفاده شده در این مقاله، ویژگیهای کپسترال حاصل از بانک فیلتر بر اساس معیار مل یعنی MFCC و ویژگیهای کپسترال حاصل از آنالیز پیشگویی خطی یا LPCC می باشند. برای استخراج پارامترهای MFCC و LPCC، سیگنال گفتار به فریم‌های ۳۵ میلی ثانیه که شروع فریم‌ها با هم ۱۰ میلی ثانیه فاصله دارند تقسیم می‌شود. پس از این بر روی سیگنال هر فریم عمل پیش تأکید با $\alpha = 0.975$ انجام می‌شود و سپس پنجرهٔ همینگ اعمال می‌گردد. در آنالیز بانک فیلتر، ۱۸ فیلتر مثلثی که بر روی طیف فرکانسی بر اساس معیار مل توزیع شده‌اند استفاده می‌شود و سپس ۱۲ ضریب کپسترال استخراج می‌شود.

برای استخراج پارامترهای LPCC، پس از اعمال پنجرهٔ همینگ، آنالیز پیشگویی خطی بروش اتوکورولیشن با مرتبهٔ P=12 انجام می‌شود و سپس از ضرایب پیشگویی خطی بدست آمده، ۱۲ ضریب کپسترال استخراج می‌گردد. بر روی پارامترهای کپسترال LPCC و MFCC، یک لیفتر کاهنده در طرفین (لیفتر جوانگ) اعمال می‌شود و سپس پارامترهای کپسترال وزن دهی شده بدست می‌آیند. آنالیز بانک فیلتر با معیار مل برای بدست آوردن ضرائب MFCC توسط ۱۸ فیلتر مثلثی صورت میگیرد. مشتق اول و دوم ضرایب کپسترال و نیز لگاریتم انرژی و مشتق اول و دوم لگاریتم انرژی نیز به ضرایب کپسترال اضافه می‌شوند. در نهایت بردارهای ویژگی، ۳۹ بعدی حاصل برای مدل نمودن اعداد بگونه‌ای که در بخشهای بعد توضیح داده خواهد شد مورد استفاده قرار گرفت.

۵ـ تفاضل میانگین در حوزه کپسترال [11]

برای بهسازی ضرائب کپسترال و کاهش تاثیر خصوصیات کانال مخابراتی، از روش تفاضل میانگین در حوزه کپسترال استفاده نموده‌ایم. اگر سیگنال گفتار را s(n)، پاسخ فرکانسی کانال انتقال را g(n)، سیگنال دریافتی پس از عبور از کانال را t(n) و ضرایب کپسترال متناظر را $c_s(n)$، $c_g(n)$ و $c_t(n)$ بنامیم، داریم:

$$T(z) = S(z).G(z) \qquad (۷)$$

$$LogT(z) = LogS(z) + LogG(z) \qquad (۸)$$

$$c_t(n) = c_s(n) + c_g(n) \qquad (۹)$$

$$c_t(m,n) = c_s(m,n) + c_g(m,n) \qquad (۱۰)$$

که m شماره فریم می‌باشد. اگر مشخصه فرکانسی کانال را در طول زمان ثابت فرض کنیم یعنی $c_g(m,n) = c_g(n)$ و نیز فرض کنیم در یک قطعه گفتار بیان شده که از لحاظ فونتیکی متعادل است، جمع پارامترهای کپسترال در طی زمان صفر می‌باشد آنگاه:

$$\sum_m c_t(m,n) = \sum_m c_s(m,n) + \sum_m c_g(m,n) \qquad (۱۱)$$

$$\sum_m c_t(m,n) = 0 + c_g(n) \qquad (۱۲)$$

معنای معادله فوق این است که میانگین‌گیری بر روی بردارهای کپسترال در طی زمان تخمین خوبی از $c_g(n)$ یعنی مشخصه کانال در حوزه کپسترال را بدست می‌دهد. پس پارامترهای کپسترال بهسازی شده بصورت زیر بدست می‌آیند:

$$c_t^n(m,n) = c_t(m,n) - \sum_{m'} c_t(m',n) \approx c_s(m,n) \qquad (۱۳)$$

یعنی با این کار تأثیر $c_g(n)$ یا مشخصه کانال از پارامترهای کپسترال حذف شده است. به این عمل، تفاضل میانگین در حوزه کپسترال یا CMS می‌گویند که برای جبران اثر کانال انتقال بر روی پارامترهای کپسترال بکار می‌رود.

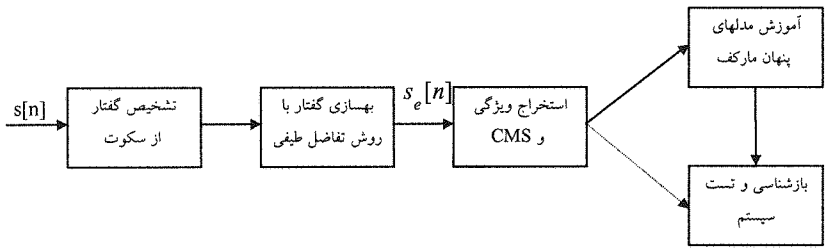
۶ـ مدل مخفی مارکف

مدل مخفی مارکف یک ماشین حالت متناهی است که از دو فرآیند اتفاقی همروند تشکیل شده است. یک فرآیند اتفاقی، توالی حالاتی است که مدل طی می‌کند و این توالی مخفی است. فرآیند اتفاقی دیگر تولید مشاهدات در هر حالت می‌باشد [13] ,[12]. بسته به نحوهٔ تولید مشاهدات دو نوع مدل وجود دارد. یک نوع آن مدل مخفی مارکف گسسته است که در آن مشاهدات، محدود به یک مجموعه الفبای خاص و گسسته بنام کتاب کد می‌باشد [12]. بنابراین باید بردارهای ویژگی به تعداد محدودی چندی شوند که این امر کاهش ذاتی دقت را به دنبال خواهد داشت. نوع دیگر، مدل مخفی مارکف از نوع پیوسته می‌باشد که با استفاده از یک تابع چگالی احتمال پیوسته از نوع مخلوط گوسی، بردارهای مشاهده را تولید می‌کند. بنابراین نیازی به عمل چندی‌سازی وجود ندارد. در این تحقیق، مدل مخفی مارکف پیوسته مورد استفاده قرار گرفته است.

۷ـ بازشناسی ارقام گسسته

۷ـ۱ـ اجزای سیستم بازشناسی ارقام گسسته

اجزای سیستم بازشناسی ارقام گسسته که در این تحقیق پیاده سازی شده است در شکل زیر مشاهده میشود. این سیستم شامل بخش تشخیص گفتار از سکوت به منظور تعیین محدوده ارقام و استفاده از بخش سکوت برای تخمین نویز زمینه، بخش بهسازی گفتار بروش تفاضل طیفی به منظور کاهش تأثیر نویز جمعی ناشی از تلفن و خطوط انتقال، بخش استخراج ویژگی و نرمالیزه کردن ضرائب کپسترال بروش تفاضل میانگین در حوزه کپسترال و نهایتا آموزش مدلهای ارقام بروش مدل مخفی مارکف و نیز بخش بازشناسی است که از الگوریتم ویتربی استفاده می‌نماید.

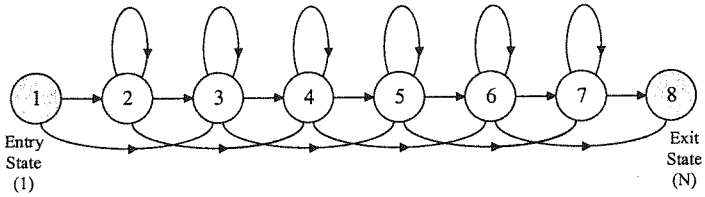


شکل (۳) اجزای سیستم بازشناسی ارقام گسسته.

بجز بخش آموزش و بازشناسی سایر بخشها در قسمت‌های قبل توضیح داده شد. در بخش‌های بعدی به توضیح چگونگی آموزش مدلهای ارقام بروش مدل مخفی مارکف و نیز چگونگی بازشناسی می‌پردازیم.

۷ـ۲ـ آموزش مدل‌های ارقام

به منظور بازشناسی ارقام گسسته، هر رقم را توسط یک مدل مخفی مارکف مدل می‌نماییم. برای این منظور از یک مدل چپ براست با ۶ حالت و تعدادی از توابع گوسی در هر حالت استفاده نمودیم. برای سکوت زمینه هم یک مدل یک حالته در نظر گرفته شد.



شکل (٤) مدل پنهان مارکف برای ارقام گسسته و متصل.

یک مدل شش حالته به اضافهٔ دو حالت ورودی و خروجی به صورت شکل ۴ خواهد بود. به منظور آموزش مدلها و نزدیک شدن به ماکزیمم‌های سراسری در تخمین پارامترهای مدل، آموزش مدل‌های مارکف در دو مرحله به شرح زیر انجام گردید:

الف ـ مرحله اول

در این مرحله برای ایجاد مدل اولیه هر رقم ابتدا به ماتریس گذر بین حالات مقادیر منطقی نسبت داده می‌شود. در اولین تکرار ابتدا کلیه بردارهای آموزشی حاصل از هر تکرار از رقم مورد نظر بطور مساوی بین حالات تقسیم شده و سپس مقادیر اولیهٔ ماتریس‌های کوواریانس و بردارهای میانگین محاسبه می‌شوند. در اینجا ماتریس‌های کوواریانس قطری هستند. در تکرارهای بعدی با استفاده از این مدل اولیه و استفاده از الگوریتم ویتربی، بهترین تخصیص بردارها به حالات مدل صورت گرفته و سپس عملیات خوشه‌بندی توسط الگوریتم K-means انجام می‌شود و پارامترهای مدل تصحیح می‌گردند تا مدل جدیدی ایجاد شود. این کار تا رسیدن به همگرایی ادامه می‌یابد.

ب ـ مرحله دوم

در این قسمت مدل حاصل از مرحله اول دریافت شده و بطور تکراری و با استفاده از فرمولهای تخمین بام ـ ولش، تا رسیدن به همگرایی، پارامترهای مدل تصحیح می‌شوند. مراحل انجام کار بدین صورت است که اگر فرض کنیم O^r نمونهٔ r ام از رقم مورد نظر باشد، متغیرهای پیشرو و پسرو یعنی $\alpha_i^r(t)$ و $\beta_i^r(t)$ را بصورت زیر تعریف کنیم:

$$\alpha_i^r(t) = P(o_1^r, o_2^r, \ldots, o_t^r, q_t = i \mid \lambda) \qquad (۱۴)$$

$$\beta_i^r(t) = P(o_{t+1}^r, o_{t+2}^r, \ldots, o_{T_r}^r \mid q_t = i, \lambda) \qquad (۱۵)$$

و نیز داشته باشیم $P_r = (O^r \mid \lambda)$ و $O^r = \{o_1^r, o_2^r, \ldots, o_t^r, \ldots, o_{T_r}^r\}$، آنگاه فرمولهای برگشتی برای محاسبهٔ $\alpha_i^r(t)$ و $\beta_i^r(t)$ بصورت زیر خواهد بود:

$$\alpha_j^r(t) = \left[\sum_{i=2}^{N-1} \alpha_i^r(t-1) a_{ij}\right] b_j(o_t^r) \qquad (۱۶)$$

$$\alpha_1^r(1) = 1 \ , \ \alpha_j^r(1) = a_{1j} b_j(o_1^r) \qquad (۱۷)$$

$$\alpha_N^r(T) = \sum_{i=2}^{N-1} \alpha_i^r(T).a_{iN} \qquad (۱۸)$$

$$\beta_i^r(t) = \sum_{j=2}^{N-1} a_{ij} b_j(o_{t+1}^r)\beta_j^r(t+1) \qquad (۱۹)$$

$$\beta_i^r(T) = a_{iN} \ , \ \beta_i^r(1) = \sum_{j=2}^{N-1} a_{1j} b_j(o_1^r)\beta_j^r(1) \qquad (۲۰)$$

یک مدل شش حالته به اضافهٔ دو حالت ورودی و خروجی به صورت شکل ۴ خواهد بود. فرمولهای تخمین بام ـ ولش بصورت زیر می‌باشند:

$$\overline{a}_{ij} = \frac{\sum_{r=1}^{R} \frac{1}{P_r} \sum_{t=1}^{T_r-1} \alpha_i^r(t) a_{ij} b_j(o_{t+1}^r) \beta_j^r(t+1)}{\sum_{r=1}^{R} \frac{1}{P_r} \sum_{t=1}^{T_r} \alpha_i^r(t) \beta_i^r(t)} \qquad (۲۱)$$

احتمال گذر از حالت ورودی به حالت‌های دیگر و نیز احتمال گذر از حالات مدل به حالت خروجی چنین تخمین زده می‌شود، حالت‌های ورودی و خروجی چون غیر مولد هستند، تابع چگالی احتمال نخواهند داشت:

$$\overline{a}_{1j} = \frac{1}{R} \sum_{r=1}^{R} \frac{1}{P_r} \alpha_j^r(1) \beta_j^r(1) \qquad (۲۲)$$

$$\overline{a}_{iN} = \frac{\sum_{r=1}^{R} \frac{1}{P_r} \alpha_i^r(T) \beta_i^r(T)}{\sum_{r=1}^{R} \frac{1}{P_r} \sum_{t=1}^{T_r} \alpha_i^r(t) \beta_i^r(t)} \qquad (۲۳)$$

$$L_{jm}^r(t) = \frac{1}{P_r} U_j^r(t) c_{jm} b_{jm}(o_t^r) \beta_j^r(t) b_j(o_t^r) \qquad (۲۴)$$

$$U_j^r(t) = \begin{cases} a_{1j} & \text{if } t = 1 \\ \sum_{i=2}^{N-1} \alpha_i^r(t-1) a_{ij} & \text{Otherwse} \end{cases} \qquad (۲۵)$$

$$\overline{\mu}_{jm} = \frac{\sum_{r=1}^{R} \sum_{t=1}^{T_r} L_{jm}^r(t) o_t^r}{\sum_{r=1}^{R} \sum_{t=1}^{T_r} L_{jm}^r(t)} \qquad (۲۶)$$

$$\overline{\Sigma}_{jm} = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T_r} L_{jm}^{r}(t)(o_t^r - \mu_{jm})(o_t^r - \mu_{jm})^T}{\sum_{r=1}^{R}\sum_{t=1}^{T_r} L_{jm}^{r}(t)} \qquad (۲۷)$$

$$\overline{c}_{jm} = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T_r} L_{jm}^{r}(t)}{\sum_{r=1}^{R}\sum_{t=1}^{T_r} L_{j}^{r}(t)} \quad , \quad L_j^r(t) = \frac{1}{P_r}\alpha_j^r(t)\beta_j^r(t) \qquad (۲۸)$$

پس از بکار بردن این فرمولها در هر مرحله، بطور تکراری و تا رسیدن به همگرایی این کار را ادامه می‌دهیم.

۷ـ۳ـ نتایج آزمایشات

برای ارقام صفر تا نه، مدلهای چپ به راست ۶ حالته که هر یک از حالت ها ۱۶، ۳۲ یا ۶۴ تابع گاوسی در خود داشتند و برای سکوت زمینه هم یک مدل یک حالته با ۳۲ تابع گاوسی در نظر گرفته شد. مدل‌ها با الگوریتم بخش۸-۲ آموزش داده شدند. پایگاه صدای استفاده شده پایگاه تلفنی FARSDIGITS1 می‌باشد که در آن هم مکالمات درون شهری و هم مکالمات برون شهری (راه دور) ضبط شده است و دارای کیفیت SNR=8.8dB می باشد. گفتارهای تلفنی ضبط شده، با استفاده از روش تفاضل طیفی و روش تشخیص گفتار از سکوت بهسازی گردیدند. پایگاه گفتار FARSDIGITS1 شامل گفتارهای ۱۰۰ گوینده است که ۶۱ نفر آنها مذکر و ۳۹ نفر آنها مؤنث هستند. برای آموزش هر رقم، ۵۴۶ نمونه از هر رقم که مربوط به ۵۰ گویندهٔ اول پایگاه داده می باشند، استفاده شده است. در مرحله بازشناسی ۶۷۶ نمونه از هر رقم که مربوط به ۵۰ گوینده دوم پایگاه داده بوده و سیستم در مرحله آموزش گفتار آنها را تجربه نکرده است استفاده گردید. بنابراین بازشناسی کاملاً مستقل از گوینده می‌باشد. نتایج بدست آمده در جدول شماره ۱ نشان داده شده‌اند.

جدول (۱) درصد صحت بازشناسی ارقام گسسته.

ویژگی	LPCC+log E+Δ+ΔΔ			MFCC+log E+Δ+ΔΔ		
	۱۶	۳۲	۶۴	۱۶	۳۲	۶۴
نمونه‌های آموزشی	۹۷/۹	۹۹/۸	۹۲/۹	۹۹/۴	۹۹/۹	۹۳/۵
نمونه‌های آزمایشی	۹۵/۲	۹۶/۸	۹۱/۷	۹۷/۷	۹۸/۶	۹۳/۴

ویژگیهای استفاده شده، ضرایب کپسترال، لگاریتم انرژی، مشتق اول و دوم ضرایب کپسترال و مشتق اول و دوم لگاریتم انرژی هستند. همانطور که مشاهده می‌شود ۳۲ تابع گوسی برای مدل کردن تابع چگالی احتمال نتایج بهتری نسبت به زمانی که تعداد توابع گوسی ۱۶ و ۶۴ هستند ارائه می‌دهد. درصد شناسایی بالای ۹۸/۶٪ به ازای نمونه‌های آزمایشی برای پایگاه دادهٔ تلفنی با SNR=8.8dB گویای کارایی خوب روش بکار رفته در بازشناسی ارقام فارسی می‌باشد. همچنین در یک آزمایش به جای مدل کردن سکوت از جداسازی گفتار از سکوت استفاده شد و مشاهده گردید که راندمان سیستم حدود ۳٪ کاهش پیدا کرد. با توجه به این موضوع تصمیم گرفته شد که سکوت با استفاده از یک مدل مارکف یک حالته مدل شود. ملاحظه می‌شود که کارایی پارامترهای کپسترال حاصل از بانک فیلتر مبتنی بر معیار مل یعنی ضرائب MFCC بهتر از پارامترهای کپسترال حاصل از آنالیز پیشگویی خطی یعنی ضرائب LPCC است. دلیل این برتری تأثیرپذیری ناچیز خروجی فیلترهای بانک فیلتر و نیز تأثیرپذیری شدید پارامترهای پیشگویی خطی از باقیمانده نویز موجود در گفتار تلفنی یا نویز موزیکال ناشی از روش حذف نویز تفاضل طیفی و نیز توانایی بیشتر ضرایب کپسترال MFCC که از خواص شنیداری گوش انسان الهام گرفته است در ارائه اطلاعات موجود در گفتار و در نتیجه تمایز بهتر اطلاعات مربوط به ارقام زبان فارسی، می‌باشد. این پدیده در مرجع [19] نیز گزارش شده است. در آزمایش دیگری که انجام شد برای جبران مشخصه کانال تلفنی از روش تفاضل میانگین در حوزه کپسترال استفاده گردید که کارایی سیستم را تا ۹۹/۱٪ افزایش داد.

۸ـ بازشناسی ارقام متصل

مراحل بازشناسی ارقام متصل با مراحل بازشناسی ارقام گسسته شکل ۳ مشابه می‌باشد. در بازشناسی کلمات متصل می‌توان سه راه در پیش گرفت: الف ـ برای بازشناسی کلمات متصل از مدل‌های بدست آمده برای بازشناسی کلمات گسسته استفاده کرد که داده‌های آموزشی این مدلها کلمات گسسته بوده باشد. ب ـ برای بازشناسی کلمات متصل از مدل‌هایی استفاده کرد که بر روی رشته کلمات متصل و با استفاده از فرمولهای تخمین بام ـ ولش آموزش دیده‌اند. ج ـ برای بازشناسی کلمات متصل از مدل‌هایی استفاده کرد که داده‌های آموزشی آنها کلمات متصل است و ابتدا توسط الگوریتم بام ـ ولش تخمین زده شده‌اند و سپس پارامترهای این مدلها بروش بازتخمین نهفته (Embedded Reestimation) تخمین مجدد زده شده باشند. از روش بازتخمین نهفته برای تخمین دوبارهٔ پارامترهای تمام مدل‌ها بطور موازی و نیز افزایش کارایی آنها استفاده گردیده است. در این مقاله از روند (ج) برای آموزش مدل‌ها استفاده گردیده است. طبیعی است که روند (ب) بهتر از روند (الف) می‌باشد. آزمایشها نشان می‌دهند که کارایی روش (ج) نیز بهتر از روش (ب) بوده وکارائی سیستم را بهبود می‌بخشد. روش بازتخمین نهفته در بخش بعدی توضیح داده خواهد شد.

۸ـ۱ـ آموزش مدل‌های ارقام

بطور خلاصه آموزش مدل های پنهان مارکف برای بازشناسی ارقام متصل طی ۴ مرحله صورت گرفت: الف ـ تقسیم یکسان بردارها بین حالات و محاسبه مقادیر اولیه بردارهای میانگین و ماتریس های کوواریانس (قطری) ب ـ تقسیم بهینه بردارهای هر تکرار از رقم مورد نظر بین حالات مدل با استفاده از الگوریتم ویتربی و تصحیح پارامترهای مدل با استفاده از خوشه‌بندی K-means و تکرار این کار تا رسیدن به همگرایی ج ـ تصحیح پارامترهای مدل بطور تکراری با استفاده از فرمولهای تخمین بام ـ ولش تا رسیدن به همگرایی دـ تصحیح پارامترهای مدل با استفاده از فرمولهای بازتخمین نهفته [13] و تکرار این کار به دفعات لازم طوری که باعث آموزش بیش از حد نشود.

تصحیح مجدد پارامترهای مدل توسط الگوریتم بازتخمین نهفته تفاوتی که با مراحل ب و ج دارد این است که پارامترهای تمام مدل‌ها بطور همزمان با هم تخمین زده می‌شوند. اگر در یک رشته از ارقام در دورهٔ آموزش، Q رقم به هم چسبیده باشند، و اگر رقم فعلی در رشته، مربوط به مدل q ام باشد، رقم بعد از آن را q+1 می‌نامیم. محاسبه احتمالات پیشرو و پسرو برای مدل q ام در زمان t=1 چنین خواهد بود:

$$\alpha_1^{(q)}(1)=\begin{cases}1 & \text{if } q=1\\ \alpha_1^{(q-1)}(1)a_{1N_{q-1}}^{(q-1)} & \text{Otherwise}\end{cases} \qquad (۲۹)$$

$$\alpha_j^{(q)}(1)=a_{1j}^{(q)}b_j^{(q)}(o_1) \qquad (۳۰)$$

$$\alpha_{N_q}^{(q)}(1)=\sum_{i=2}^{N_q-1}\alpha_i^{(q)}(1)a_{iN_q}^{(q)} \qquad (۳۱)$$

که در آن q درون پرانتز نشاندهندهٔ اندیس مدل کلمه و N_q نشاندهندهٔ تعداد حالات مدل q ام می‌باشد. برای لحظات $t>1$، داریم:

$$\alpha_1^{(q)}(t)=\begin{cases}0 & \text{if } q=1\\ \alpha_{N_{q-1}}^{(q-1)}(t-1)+\alpha_1^{(q-1)}(t)a_{1N_{q-1}}^{(q-1)} & \text{Otherwise}\end{cases} \qquad (۳۲)$$

$$\alpha_j^{(q)}(t)=\left[\alpha_1^{(q)}(t)a_{1j}^{(q)}+\sum_{i=2}^{N_q-1}\alpha_i^{(q)}(t-1)a_{ij}^{(q)}\right]b_j^{(q)}(o_t) \qquad (۳۳)$$

$$\alpha_{N_q}^{(q)}(t)=\sum_{i=2}^{N_q-1}\alpha_i^{(q)}(t)a_{iN_q}^{(q)} \qquad (۳۴)$$

برای بدست آوردن احتمالات پسرو $\beta_i(t)$ چنین عمل می کنیم، اگر t=T باشد:

$$\beta_{N_q}^{(q)}(1)=\begin{cases}1 & \text{if } q=Q\\ \beta_{N_{q+1}}^{(q+1)}(T)a_{1N_{q+1}}^{(q+1)} & \text{Otherwise}\end{cases} \qquad (۳۵)$$

$$\beta_i^{(q)}(T)=a_{iN_q}^{(q)}\beta_{N_q}^{(q)}(T) \qquad (۳۶)$$

$$\beta_1^{(q)}(T)=\sum_{j=2}^{N_q-1}a_{1j}^{(q)}b_j^{(q)}(o_T)\beta_j^{(q)}(T) \qquad (۳۷)$$

برای $t\prec T$ داریم:

$$\beta_{N_q}^{(q)}(t)=\begin{cases}0 & \text{if } q=Q\\ \beta_1^{(q+1)}(t+1)+\beta_{N_{q+1}}^{(q+1)}(t)a_{1N_{q+1}}^{(q+1)} & \text{Otherwise}\end{cases} \qquad (۳۸)$$

$$\beta_i^{(q)}(t)=a_{iN_q}^{(q)}\beta_{N_q}^{(q)}(t)+\sum_{j=2}^{N_q-1}a_{ij}^{(q)}b_j^{(q)}(o_{t+1})\beta_j^{(q)}(t+1) \qquad (۳۹)$$

$$\beta_1^{(q)}(t)=\sum_{j=2}^{N_q-1}a_{1j}^{(q)}b_j^{(q)}(o_t)\beta_j^{(q)}(t) \qquad (۴۰)$$

احتمال مشاهدات روی مدل q ام یعنی $P(O|\lambda_q)$ به این طریق بدست می‌آید:

$$P(O|\lambda_q)=\alpha_{N_q}^{(q)}(T)=\beta_1^{(q)}(1) \qquad (۴۱)$$

فرمولهای تخمین بام ـ ولش در حالت باز تخمین نهفته متفاوت بوده و بگونه‌ای تغییر می‌کند که احتمال گذر از حالت ورودی به حالت‌های دیگر در مدل q ام بصورت زیر تخمین زده می‌شود:

$$\overline{a}_{1j}^{(q)}=\frac{\sum_{r=1}^{R}\frac{1}{P_r}\sum_{t=1}^{T_r-1}\alpha_1^{(q)r}(t)a_{1j}^{(q)}b_j^{(q)}(o_t^r)\beta_j^{(q)r}(t)}{\sum_{r=1}^{R}\frac{1}{P_r}\sum_{t=1}^{T_r}\alpha_i^{(q)r}(t)\beta_i^{(q)r}(t)+\alpha_1^{(q)r}(t)a_{1N_q}^{(q)}\beta_1^{(q+1)r}(t)} \qquad (۴۲)$$

احتمال گذر از حالت‌های داخلی به حالت خروجی از رابطه زیر بدست می‌آید:

$$\overline{a}_{iN_q}^{(q)} = \frac{\sum_{r=1}^{R}\frac{1}{P_r}\sum_{t=1}^{T_r-1}\alpha_i^{(q)r}(t)a_{iN_q}^{(q)}\beta_{N_q}^{(q)r}(t)}{\sum_{r=1}^{R}\frac{1}{P_r}\sum_{t=1}^{T_r}\alpha_i^{(q)r}(t)\beta_i^{(q)r}(t)} \qquad (۴۳)$$

احتمال گذر از حالت ورودی به حالت خروجی بصورت زیر محاسبه می شود:

$$\overline{a}_{1N_q}^{(q)} = \frac{\sum_{r=1}^{R}\frac{1}{P_r}\sum_{t=1}^{T_r-1}\alpha_1^{(q)r}(t)a_{1N_q}^{(q)}\beta_1^{(q+1)r}(t)}{\sum_{r=1}^{R}\frac{1}{P_r}\sum_{t=1}^{T_r}\alpha_i^{(q)r}(t)\beta_i^{(q)r}(t)+\alpha_1^{(q)r}(t)a_{1N_q}^{(q)}\beta_1^{(q+1)r}(t)} \qquad (۴۴)$$

۸ـ۲ـ نتایج آزمایشات

در این سیستم، برای آموزش مدل‌های ارقام در حالت متصل، از گفتار۳۰ نفر از گویندگان (مذکر و مؤنث) پایگاه صدای تلفنی FARSDIGITS1 [5] استفاده شده است. گفتیم که در این پایگاه صدا، هر گوینده ۱۰۰ رشته دو رقمی متصل (تمام حالات ممکن) را بیان نموده است. برای آموزش مدل‌ها از تمام نمونه‌های آن رقم متعلق به ۱۵ گوینده اول استفاده گردید و مدل‌ها طبق مراحل الف، ب، ج و د آموزش داده شدند و برای بازشناسی از نمونه‌های رشته‌های ارقام ادا شده متعلق به ۱۵ گوینده دوم استفاده گردید. با توجه به تلفنی بودن پایگاه دادهٔ مورد نظر و برای مقابله با نویز از الگوریتم مقاوم برای تشخیص گفتار از سکوت (بخش ۳) و نیز از روش تقاضل طیفی (بخش ۴) برای بهسازی گفتار استفاده گردید. نتایج حاصل از آزمایشات در جدول زیرآمده است. این نتایج به ازای پارامترهای مبتنی بر بانک فیلتر توزیع شده بر اساس معیار مدل، بدست آمده‌اند.

همانگونه که در جدول مشاهده می‌شود در صورت عدم استفاده از روش بازتخمین نهفته صحت بازشناسی ارقام متصل ۱/ ۸۳٪ می‌باشد. این کارایی در صورت استفاده از روش باز تخمین نهفته تا ۸۳/۷٪ افزایش می‌یابد. بیشترین افزایش بازای ۳ بار تکرار تخمین بدست آمده است. نتایج نشان می‌دهد که اگر تعداد دفعات بکار بردن این تخمین زیاد شود، به علت آموزش بیش از حد مدل‌ها، راندمان برای داده‌های آزمایشی افت می‌نماید.

جدول (۲) درصد صحت بازشناسی ارقام متصل.

داده های بازشناسی	آموزش بدون بازتخمین نهفته	تعداد دفعات تخمین مجدد در آموزش بروش بازتخمین نهفته					
		۱	۲	۳	۴	۵	۶
آموزشی	۹۰/۱	۹۰/۷	۹۱/۱	**۹۱/۹**	۹۱/۳	۹۰/۹	۹۰/۳
آزمایشی	۸۳/۱	۸۳/۵	۸۳/۷	**۸۳/۷**	۸۳/۶	۸۳/۵	۸۳/۵

۹ـ بازشناسی اعداد پیوسته

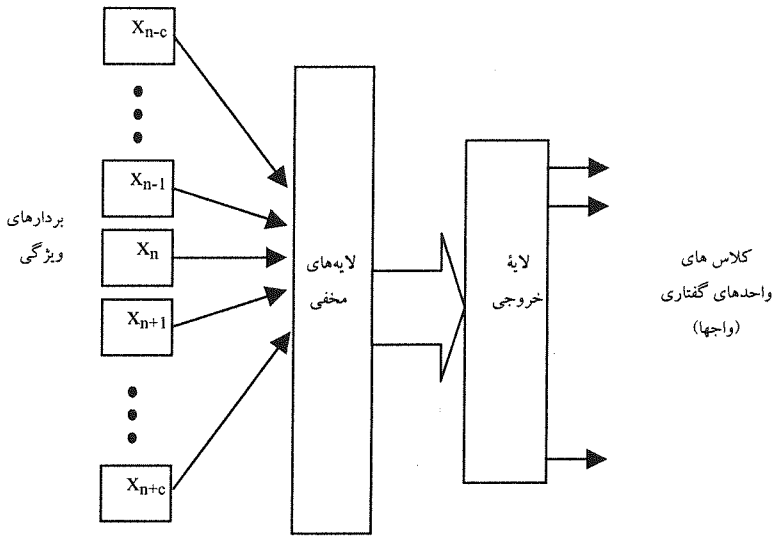
دو بخش قبل به بازشناسی اعداد و ارقام گسسته و متصل اختصاص داشت. در این بخش به توصیف آزمایشات صورت گرفته برای بازشناسی اعداد پیوسته می‌پردازیم. در اینجا تعداد ارقام هر عدد نامحدود بوده و نحوهٔ خواندن آن به گونه‌های مختلف امکان‌پذیر است. بعنوان مثال، فرض کنید که در اینجا یک عدد ۸ رقمی مورد نظر باشد. بدیهی است که یک عدد ۸ رقمی را می‌توان به شکل‌های گوناگون بیان کرد. مثلاً عدد ۱۲۳۴۵۶۷۸ را در نظر بگیرید. یک راه برای بیان این عدد خواندن آن بصورت یکپارچه است که نتیجه آن عبارتست از دوازده میلیون و سیصد و چهل و پنج هزار و شش‌صد و هفتاد و هشت. یک راه دیگر این است که ارقام آنرا یکی‌یکی به شکل زیر بخوانیم: یک، دو، سه، چهار، پنج، شش، هفت، هشت. روش دیگر این است که آنرا به اعداد دو رقمی تقسیم کرده و سپس آن اعداد را بیان نمائیم مانند: دوازده، سی و چهار، پنجاه و شش، هفتاد و هشت. همچنین می‌توان آنرا بصورت ترکیبی از اعداد دو رقمی و سه رقمی بیان کرد: دوازده، سیصد و چهل و پنج، شش‌صد و هفتاد و هشت و ... . می‌بینیم که روش‌های بیان کردن اعداد می‌تواند بسیار متنوع باشد و اگر محدودیتی وجود نداشته باشد، مسأله پیچیده خواهد شد. بنابراین در بازشناسی اعداد پیوسته، گوینده هیچ محدودیتی در نحوهٔ بیان اعداد ندارد و به هر شکلی که مایل است می‌تواند آنها را بیان نماید. بدلیل پیوسته بودن نوع بازشناسی، واحد صوتیِ مناسب برای این منظور واج می‌باشد.

در اینجا نیز واحد صوتی پایه در بازشناسی اعداد پیوسته واج انتخاب گردید. از سه روش مدل مخفی مارکف، شبکه عصبی و نیز سیستم هیبرید ناشی از ترکیب آنها برای مدل نمودن واجها استفاده گردید.

۹ـ۱ـ مدل مخفی مارکف

برای هر واج یک مدل مخفی مارکف ایجاد گردید. واج /?/ (که بعنوان مثال در کلمهٔ "ایزد" /izad?/ وجود دارد) در مجموعهٔ اعداد فارسی بکار گرفته نمیشود. بنابراین برای آن مدلی ساخته نشد. هر مدل دارای ۳ حالت میباشد. برای هر حالت، احتمال خروج از آن حالت ($a_{i\,Out}$) نیز در نظر گرفته شد که احتمال گذر از آن حالت به حالت ورودی مدل دیگر را نشان می‌دهد. تعداد توابع گاوسی برای تمامی مدل‌ها یکسان و برابر با ۵ قرار داده شد.

مقداردهی اولیهٔ پارامترهای مدل، با استفاده از روش بخش‌بندی یکنواخت انجام گرفت و پارامترهای توابع گاوسی هر حالت نیز با استفاده از الگوریتم خوشه‌بندیSplitting LBG بدست آمد. در این الگوریتم تغییر کوچکی اعمال گردید. اگر تقسیم هر بردار به دو بخش موجب گردد که تعداد کل خوشه‌ها بیش از تعداد مورد نظر شود، فقط خوشهٔ دارای حداکثر اعوجاج به دو بخش تقسیم می‌شود. در آموزش مدل، الگوریتم آموزش ویتربی مورد استفاده قرار گرفت. در مرحلهٔ بازشناسی احتمال گذر از یک مدل به مدل دیگر به طور یکسان برابر $\frac{1}{29}$ قرار داده شد و هیچ نوع گرامری اعمال نگردید



شکل (۵) معماری سیستم بازشناسی مبتنی بر MLP برای بازشناسی واجها.

۹ـ۲ـ شبکهٔ عصبی

شبکه‌های عصبی مصنوعی روشی برای کاهش وابستگی سیستم به فرضیات غیرواقعی دربارهٔ گفتار می‌باشد [14, 15]. انواع متعددی از این شبکه‌ها در کاربردهای بازشناسی مورد استفاده قرار گرفته‌اند که به عنوان مثال می‌توان از پرسپترون چندلایهMLP. نام برد. در این تحقیق نیز از شبکهٔ عصبی MLP استفاده شده است. آموزش این شبکه‌ها، *تمایزی* (Discriminative) است، یعنی در خلال آموزش سعی می‌شود شباهت الگو به کلاسی که در آن قرار گرفته است حداکثر شده و در عین حال، تفاوت میان الگو و سایر کلاس‌ها تا حد ممکن افزایش یابد. به فرض مستقل بودن آماری بردارهای ویژگی، که در مدل HMM انجام می‌گرفت، در اینجا نیازی نخواهد بود. علاوه بر این، نیازی نیست فرض نماییم که تابع چگالی احتمال مشاهدات، ترکیب خطی از چند تابع گاوسی است. شبکه عصبی شکل پیچیده توزیع بردارهای ویژگی در فضای چند بعدی را با استفاده از مجموعه‌ای از ابر صفحات جدا کننده، مدل می نماید. ساختار توزیع شده و موازی این شبکه مزیت دیگر آن است، زیرا پیاده‌سازی سخت‌افزاری آن بر روی ماشینهای موازی را امکان‌پذیر می‌سازد. معماری شبکه عصبی مورد استفاده در این تحقیق در شکل زیر نشان داده شده است.

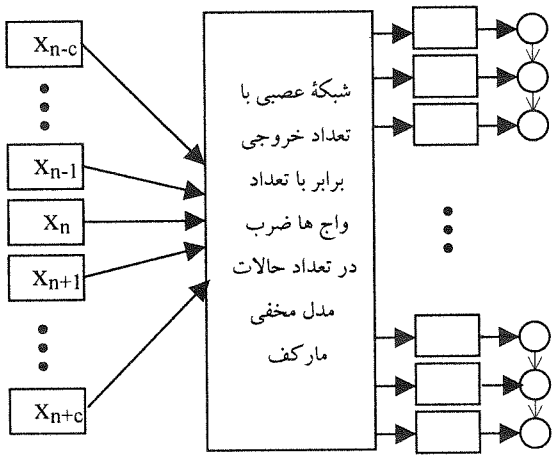
ساده‌ترین راه انجام بازشناسی با استفاده از MLP، تبدیل الگوهای زمانی به الگوهای مکانی در ورودی شبکه می‌باشد. برای

این منظور بردارهای ویژگی در زمانهای مختلف در یک همسایگی محلی بطور همزمان به شبکه اعمال میشوند. بردار ویژگیِ فعلی، c بردار قبل و c بردار بعد از آن همزمان به ورودی شبکه اعمال می‌گردد [14, 16]. در نهایت هر بردار ورودی شبکه شامل 2c+1 بردار ویژگی خواهد بود.

جهت آموزش شبکه، الگوریتم انتشار خطا به عقب تطبیقی مورد استفاده قرار گرفت. در این الگوریتم، ابتدا خطای کل بدست می‌آید. سپس با روش متداول انتشار خطا به عقب، وزن‌های جدید محاسبه می‌گردد. در پایان هر epoch خطای کل محاسبه می‌شود. اگر نسبت خطای حاصل به خطای مرحلهٔ قبل از اندازهٔ معینی (مثلاً ۱/۰۴) بیشتر باشد، وزن‌های بدست آمده پذیرفته نمی‌شود و نرخ آموزش نیز به گونه‌ای کاهنده (مثلاً ۰/۷)، کاهش می‌یابد. در غیر این صورت، وزن‌های جدید پذیرفته می‌شود و نرخ آموزش نیز به گونه‌ای افزاینده (مثلاً ۱/۰۵)، افزایش می‌یابد. شبکهٔ MLP مورد استفاده، سه لایه در نظر گرفته شد. تعداد خروجی شبکه برابر با ۲۹ نورون قرار داده شد. پارامتر c موجود در شکل ۲ برابر با ۲ انتخاب گردید. بنابراین طول زمانیِ بردارِ ویژگی ۵ فریم می‌باشد. در لایهٔ مخفی نیز ۳۰۰ نورون قرار داده شد. تابع فعالیت کلیهٔ واحدها، بطور یکسان سیگموئید انتخاب گردید.

۹ـ۳ـ ترکیب مدل مخفی مارکف و شبکهٔ عصبی

به منظور استفاده از مزایای روش‌های فوق و رفع عیوب آنها، روش‌های ترکیبی مورد استفاده قرار گرفته‌اند [14,16,17]. این روش‌ها به دو دستهٔ کلی سری و موازی تقسیم می‌شوند که در ترکیب موازی چند سیستم مستقل مورد استفاده قرار می‌گیرند و هر کدام جداگانه آموزش می‌یابند، ولی در ترکیب سری، خروجی یک سیستم ساده، ورودی برای یک سیستم سادهٔ دیگر می‌باشد. آموزش این سیستم‌ها به موازات هم انجام می‌گیرد [15].



شکل (٦) معماری سیستم ترکیبی.

در این مقاله، ترکیب مدل مخفی مارکف و شبکهٔ عصبی MLP مورد بررسی قرار گرفته است. در اینجا از شبکه عصبی MLP برای محاسبه احتمال تولید یک بردار مشاهده در یک حالت از مدل پنهان مارکف استفاده میشود. معماری سیستم در شکل فوق مشاهده می‌گردد. متناظر با هر حالت از هر مدل مخفی مارکف، یک نورون خروجی در شبکهٔ عصبی وجود دارد که احتمال مورد نیاز الگوریتم ویتربی را تولید می‌نماید [14, 17]. ورودی‌های شبکه مانند حالتی است که شبکه بطور مستقل مورد استفاده قرار گیرد. خروجی‌های شبکه، احتمالات پسین (*a posteriori*) را تولید می‌نمایند ولی در الگوریتم ویتربی مقادیر احتمالات درستنمائی یا احتمال تولید بردار در حالت استفاده می‌شود. بنابراین باید خروجی‌های شبکه به این مقادیر تبدیل گردد. فرض کنید y_i خروجی i اُم شبکه عصبی است. اگر q_i یکی از حالت‌های یکی از مدلهای واجی مارکف و متناظر با خروجی i اُم شبکه عصبی باشد داریم:

$$y_i \approx P(q_i \mid x) \quad , \quad x = [x_{n-c} \ldots x_{n-1} x_n x_{n+1} \ldots x_{n+c}] \qquad (۴۵)$$

طبق قانون Bayes خواهیم داشت:

$$P(x \mid q_i) = \frac{P(q_i \mid x).P(x)}{P(q_i)} \approx m.\frac{y_i}{P(q_i)} \approx \frac{y_i}{P(q_i)} \qquad (۴۶)$$

در اینجا چون P(x) بازای حالت های مختلف (i های مختلف) یکسان است، آن را بصورت یک عدد ثابت m فرض می‌کنیم که در مقایسه تاثیری ندارد و نیز احتمال پیشین حالت یعنی $P(q_i)$ را با استفاده از درصد رخداد هر حالت در داده‌های آموزشی بدست می‌آوریم. بنابر این برای تبدیل خروجی شبکه عصبی به احتمال تولید بردار در حالت، خروجی شبکه را بر عدد $P(q_i)$ تقسیم می نمائیم. تعداد خروجی‌های شبکه برابر است با تعداد حالات مدل مارکف بازای هر واج (N) ضربدر تعداد واجها. خروجی شبکه عصبی نیز باید مقادیری در فاصلهٔ [0,1] داشته باشند. روش آموزش ترکیبی چنین است: ابتدا نمونه‌های آموزشی واجها بطور یکنواخت به حالات مدل مارکف تخصیص داده میشود. مقادیر اولیه ماتریس گذر حالات تخمین زده میشود و با این داده‌های تخصیص داده شده به حالات، شبکه عصبی آموزش می‌بیند تا شکل توزیع بردارهای ویژگی را با استفاده از ابر صفحات بدست آورد و تابع چگالی احتمال جدید بازاء حالات مدلهای مارکف واجی بدست آیند. پس از این کار با مشخص بودن ماتریس گذر حالات و نیز توابع چگالی احتمال، با استفاده از الگوریتم ویتربی بردارهای مشاهدات را بطور بهینه به حالات مدلهای مارکف تخصیص می‌دهیم و با معلوم شدن تخصیص مشاهدات به حالات، ماتریس گذر حالات را تخمین می‌زنیم و شبکه عصبی را نیز آموزش می‌دهیم تا تابع چگالی احتمال جدید را بدست آوریم. این قدمها را تا رسیدن به همگرائی ادامه خواهیم داد.

۹ـ٤ـ پایگاه داده

روش گفته شده بر روی پایگاه داده‌های تلفنی از مجموعهٔ اعداد فارسی بنام FARSNUMBERS1 اعمال گردید. در این پایگاه، ترکیب گسسته، متصل و پیوستهٔ کلیهٔ کلمات لازم برای پوشش همهٔ اعداد تا سقف میلیارد (شامل صفر، یک، دو . . . و میلیارد) موجود می‌باشد. این ترکیبات توسط ۵۷ گوینده شامل ۴۶ گویندهٔ مذکر و ۱۱ گویندهٔ مونث بیان گردید. گویش‌های بیان شده از طریق خط تلفن و پس از عبور از یک واسط تلفنی با فرکانس ۱۱۰۲۵ Hz نمونه‌برداری شد. نمونه‌ها با ۱۶ بیت دقت نمایش داده شد. داده‌های فوق به دو بخش آموزشی و آزمایشی تقسیم گردید. تعداد ۴۰ گوینده شامل ۳۳ گویندهٔ مذکر و ۷ گویندهٔ مونث برای نمونه‌های آموزشی در نظر گرفته شد و از هر گوینده به ازای هر واج، حداکثر ۵ تکرار استخراج گردید. باقیماندهٔ نمونه‌ها شامل ۱۳ گویندهٔ مذکر و ۴ گویندهٔ مونث به عنوان مجموعهٔ آزمایشی در نظر گرفته شد.

۹-۵ـ نتایج آزمایشات

نتایج حاصل از آزمایشات در ذیل ارائه گردیده است.

۹-۵-۱ـ بازشناسی واج

ضرایب ویژگی مورد استفاده، ضرایب LPCC و MFCC و مشتق اول آنها می‌باشد. بر روی فریم‌های به طول ۲۰ms و با هم‌پوشانی ۱۰ms، پس از پیش‌تأکید با استفاده از فیلتر $1-0.95z^{-1}$، پنجرهٔ همینگ اعمال گردید. از هر فریم ۱۰ ضریب LPC استخراج شد و سپس ضرایب LPCC با ابعاد ۱۲ بدست آمد. از هر فریم پس از پیش‌تأکید و اعمال پنجره و FFT به طول ۲۵۶ نقطه و قرار دادن ۱۸ فیلتر مثلثی بر روی طیف حاصل، ۱۲ ضریب MFCC استخراج شد. بر روی ضریب‌های حاصل لیفتر سینوسی جوانگ اعمال گردید. در جدول ۳ نتایج بازشناسی واج ارائه گردیده است. همانگونه که مشاهده می‌شود، استفاده از ضرایب MFCC نسبت به ضرایب LPCC پاسخ بهتری تولید می‌نماید. از آنجایی که سیگنال گفتاری مورد استفاده تلفنی و

نویزی می‌باشد، بنابراین ضرایب مبتنی بر بانک فیلتر در این محیطها بدلیل مقاوم بودن در مقابل نویز راندمان بهتری از خود نشان میدهد. همچنین، استفاده از مشتق ضرایب به همراه خود ضرایب نتایج بهتری بدست می‌دهد. این بدان دلیل است که ضرائب مشتق حاوی اطلاعات دینامیک گفتار می‌باشند.

جدول (۳) درصد صحت بازشناسی واجهای بکار رفته در اعداد.

	HMM		MLP		MLP/HMM	
	آموزش	آزمایش	آموزش	آزمایش	آموزش	آزمایش
LPCC+ΔLPCC	۶۸/۲	۶۴/۴	۶۶/۱	۶۴/۱	۵۵/۴	۵۱/۱
MFCC+ΔMFCC	**۷۱/۴**	**۶۸/۲**	**۶۸/۵**	**۶۵/۴**	**۵۷/۲**	**۵۴/۳**
LPCC	۶۶/۸	۶۱/۸	۶۴/۰	۶۰/۹	۵۳/۰	۵۰/۲
MFCC	۶۹/۳	۶۵/۲	۶۶/۴	۶۲/۲	۵۶/۲	۵۱/۹

همانطور که ملاحظه میگردد راندمان مدل پنهان مارکف از شبکه عصبی و نیز از مدل هیبرید بالاتر است. علت این امر میتواند عوامل زیر باشد:

۱ـ آموزش در مدل پنهان مارکف بر مبنای ریاضی قویتری مبتنی است و بطور تضمین شده در هر تکرار از آموزش سیستم رو به بهبود میرود در حالیکه در شبکه عصبی ممکن است در یک تکرار نسبت به تکرار قبل خطا افزایش یابد و تعیین مقدار نرخ یادگیری بطور بهینه بسیار مشکل و حتی غیر ممکن است و همین امر حساسیت شبکه عصبی در آموزش و نیز مدت زمان آموزش را بیشتر می‌کند.

۲ـ شبکه عصبی توزیع آماری بردارهای ویژگی در فضای چند بعدی را مدل نمی‌کند. بعنوان مثال اگر بردارهای ویژگی بازاء یک واج در ۷۵٪ مواقع در ناحیه A و در ۲۵٪ از مواقع در ناحیه B از فضای چند بعدی واقع شود شبکه عصبی این احتمال رخداد را در نظر نمیگیرد و فقط بدنبال پیدا کردن مرزهای جدا کننده کلاسها از همدیگر است. حال آنکه مدل پنهان مارکف این پدیده را بخوبی در نظر میگیرد. این پدیده خود را اینچنین نشان میدهد که سطح زیر منحنی تابع چگالی احتمال در مدل پنهان مارکف پیوسته یک است ولی در شبکه عصبی MLP سطح زیر منحنی تابع چگالی احتمال یک نیست و شبکه عصبی پراکندگی و فشردگی داده‌ها در فضای چند بعدی را مدل نمی‌کند و این یک ضعف بزرگ برای شبکه عصبی و نیز برای سیستم هیبرید است. این عیب فقط مربوط به شبکهٔ عصبی MLP نیست بلکه شبکه‌های عصبی دیگر حتی RBF نیز که در آن خروجی ترکیب خطی از توابع گاوسی است، از این ضعف رنج می‌برد و بعبارت دیگر در RBF نیز سطح زیر منحنی تابع چگالی احتمال یک نیست.

۳ـ شبکه عصبی بتنهایی بطور محلی و در حد چندین فریم زمانی قادر به مدل کردن دینامیک محلی گفتار است حال آنکه مدل پنهان مارکف دینامیک سراسری گفتار را در حد وسیع‌تری (واج، کلمه و گویش) با استفاده از برنامه‌ریزی پویا مدل می‌کند.

۴ـ در کارهای انجام شده در [15] گزارش شده است که با مساوی بودن تعداد پارامترهای دو سیستم، سیستم هیبرید بهتر از مدل پنهان مارکف پیوسته کار می‌کند ولی در حالت نابرابر بودن پارامترها، ظرفیت نهائی مدل مارکف پیوسته بیش از ظرفیت نهائی سیستم هیبرید است. این در واقع به خاطر معایب شبکه عصبی بعنوان یک تخمین‌گر برای تابع چگالی احتمال در سیستم هیبرید می‌باشد.

۵ـ راندمان مدل پنهان مارکف پیوسته برای بازشناسی واج بیشتر از شبکه عصبی است ولی برای مقایسه دقیق‌تر بین مدل پنهان مارکف پیوسته و سیستم هیبرید باید بازاء تعداد دفعات زیاد، هم مدل پنهان مارکف پیوسته و هم سیستم هیبرید MLP/HMM با شروع از نقاط اولیه تصادفی مختلف آموزش داده شوند و میانگین راندمان و یا بهترین راندمان بازاء این دفعات مختلف با هم مقایسه گردد که در کارهای انجام شده تا کنون چنین مقایسه جامعی صورت نگرفته است. این مقایسه جامع از آن جهت لازم است که هم مدل پنهان مارکف و هم شبکه عصبی در آموزش خود نسبت به مقادیر اولیه پارامترها حساس هستند. علاوه بر این آموزش هر کدام از این سیستم ها نیز می‌تواند با استفاده از تکنیک‌های بهینه‌سازی سراسری همچون سرد کردن فلزات (Simulated Annealing)، الگوریتمهای ژنتیکی و یا تکنیک‌های دیگر صورت گیرد که در اینگونه

موارد نیاز به کامپیوترهای بسیار قدرتمند وجود دارد. عقیده نویسندهٔ این مقاله، مرجع [15] و نتایج ارائه شده در آن مؤید این مطلب است که قدرت نهایی مدل پنهان مارکف پیوسته بیش از سیستم هیبرید MLP/HMM میباشد و دلیل آن ضعف شبکه عصبی در مدل کردن آماری توزیع دادهها در فضای چند بعدی است.

۶ـ بنظر میرسد قابلیت تعمیم مخلوط گاوسی در مدل پنهان مارکف برای دادههای آزمایشی و نادیده، در مدل کردنِ نحوهٔ توزیع بردارهای ویژگی در فضای چند بعدی بیش از قابلیت تعمیم مجموعهای از ابر صفحات است که توسط شبکه عصبی در مدل هیبرید بکار گرفته میشود.

۷ـ اگر در تولید واجها دو مشخصه را در نظر بگیریم، یکی طول زمانی هر یک از قطعات شبهایستا از واج (که متناظر با یکی از حالات مدل پنهان مارکف است) و دیگری نحوهٔ توزیع بردارهای ویژگی در فضای چندبعدی، آنگاه میتوان گفت: شبکهٔ عصبی بتنهایی با هیچکدام از این مشخصهها برخورد آماری و احتمالاتی نمیکند و مدل هیبرید فقط با مشخصهٔ اول برخورد آماری دارد ولی برخورد مدل پنهان مارکف با هر دو مشخصه، یک برخورد آماری و احتمالاتی است.

۹-۵-۲ـ بازشناسی کلمه

رشته برچسبهای واجی بازشناسی شده برای استخراج کلمات از گویشِ حاویِ اعدادِ بیان شده مورد استفاده قرار گرفت. برای این منظور، یک روش برنامهریزی پویا استفاده گردید. کل کلمات شناخته شدهٔ سیستم به این الگوریتم داده شد. این الگوریتم رشته برچسبهای واجی مربوط به گویش را دریافت نموده و با استفاده از کلمات فوق، بهترین رشتهٔ کلمات با حداقل هزینه (بهترین تطبیق) را استخراج میکند. این الگوریتم تا دو خطای پشت سرهم در واج های تشکیل دهند یک کلمه را تحمل میکند. حالات خاصی که در محاورات عمومی کاربرد دارد مانند کلمات هیژده، دیویست نیز در نظر گرفته شد.

جدول (٤) نتایج بازشناسی کلمه بر حسب درصد.

	HMM		MLP		MLP/HMM	
	آموزش	آزمایش	آموزش	آزمایش	آموزش	آزمایش
LPCC+ΔLPCC	۸۷/۴	۸۶/۰	۸۴/۰	۸۱/۳	۷۸/۵	۷۱/۲
MFCC+ΔMFCC	۹۴/۰	**۹۱/۱**	۸۹/۰	**۸۴/۳**	۸۰/۷	**۷۵/۰**
LPCC	۸۳/۲	۸۰/۵	۸۱/۳	۷۷/۳	۷۴/۹	۷۰/۴
MFCC	۸۹/۲	۸۷/۳	۸۵/۲	۷۹/۵	۷۹/۳	۷۲/۳

نتایج حاصل از استخراج کلماتِ تشکیل دهندهٔ اعداد در جدول ۴ مشاهده میگردد. با توجه به این جدول، بیشترین راندمان بازای ویژگیهای MFCC و مشتق آنها در حالتی است که از روش برنامهریزی پویا برای استخراج کلمات از رشتهٔ واجهای بازشناسی شده به روش HMM استفاده شده باشد. با توجه به اینکه در روش برنامهریزی پویای بکار رفته تا دو خطای متوالی در واجهای بازشناسی شدهٔ یک کلمه قابل تحمل است، بهبود چشمگیر راندمان بازشناسی کلمات نسبت به راندمان بازشناسی واجهای تشکیلدهندهٔ اعداد در مقایسهٔ بین جداول ۳ و ۴ قابل توجیه میباشد. بررسی کلی نتایج این بخش نشان میدهد که افزایش کم در نرخ بازشناسی واج موجب افزایش زیاد در نرخ بازشناسی کلمه میگردد.

۱۰ـ نتیجه گیری

در این مقاله برای بازشناسی ارقام گسسته و متصل از مدل پنهان مارکف پیوسته با مدلهای کلمهای و برای بازشناسی اعداد پیوستهٔ فارسی بصورت نامقید از مدل پنهان مارکف پیوسته، شبکهٔ عصبی و روش هیبرید با مدلهای واجی استفاده گردید. برای مقابله با نویز جمعشونده بر روی خط تلفن از روش تفاضل طیفی و برای جبران اثر کانال انتقال تلفنی و دهنیهای مختلف از روش تفاضل میانگین در حوزهٔ کپسترال استفاده شد. برای تعیین محدودهٔ گفتار و تشخیص آن از سکوت، از یک روش مقاوم در برابر نویز استفاده گردید. آزمایشها نشان دادند که در محیط نویزی و تلفنی، کاراییِ پارامترهای کپسترالِ حاصل

از آنالیزِ بانک فیلتر با توزیعِ فرکانسیِ مل بیش از کاراییِ پارامترهای کپسترالِ حاصل از پیشگویی خطی است. همچنین مشخص گردید که مدل کردنِ سکوت نتیجه‌ای بهتر از حذف سکوت با روشهای تشخیص ابتدا و انتهای گفتار می‌دهد. نتیجهٔ بدست آمده برای بازشناسیِ ارقام گسسته بازای داده‌های آزمایشی و بر روی پایگاه دادهٔ تلفنی FARSDIGITS1 با SNR=8.8dB برابر با ۹۹/۱٪ است که گویای راندمانِ نسبتاً بالای سیستمِ بکار گرفته شده در این مقاله می‌باشد. برای بازشناسیِ ارقامِ گسسته و متصل، واحدِ بازشناسی، کلمه انتخاب شد. در بازشناسیِ ارقام متصل از روشِ بازتخمینِ نهفته استفاده گردید. روش بازتخمینِ نهفته پس از مرحلهٔ تخمینِ پارامترها توسطِ الگوریتمِ بام ـ ولش بکار می‌رود و هدفِ آن تخمینِ بهترِ پارامترهای مدل در حالت‌های مرزیِ کلمه و لحاظ کردنِ تأثیرِ ارقامِ متصل بر یکدیگر است که نسبت به وقتی که فقط از روش بام ـ ولش برای آموزشِ مدلهایِ ارقام متصل استفاده شود، درصد صحتِ بازشناسی را افزایش می‌دهد. اگر تعداد تکرارهای بازتخمین نهفته کم یا بیش از حد باشد، کارایی سیستم افت خواهد کرد. با استفاده از این روش کارایی یک سیستم بازشناسی ارقام متصل با رشته‌های دو رقمی برای داده‌های آزمایشی از ۸۳/۱۵٪ به ۸۳/۷٪ افزایش یافت.

برای بازشناسی اعداد پیوستهٔ فارسی بصورت نامقید، واحدِ بازشناسی واج انتخاب گردید و از روشهای مدل پنهان مارکف، شبکهٔ عصبی و روش هیبرید استفاده شد. آزمایشها نشان داد که کاراییِ مدل پنهانِ مارکف هم از شبکهٔ عصبی MLP و هم از ترکیب مدل پنهانِ مارکف و شبکهٔ عصبی (مدلِ هیبرید) بهتر است. زمان آموزش شبکهٔ عصبی نسبت به زمانِ آموزشِ مدل پنهان مارکف بیشتر بوده ولی زمان بازشناسی در شبکهٔ عصبی کمتر از مدل پنهان مارکف است. زمان آموزش و آزمایش در سیستم ترکیبی از هر کدام از دو روش دیگر بیشتر است. برای بازشناسی اعداد پیوستهٔ فارسی، با استفاده از یک روش برنامه‌ریزیِ پویا، بهترین توالیِ کلمات از رشتهٔ برچسب‌های واجیِ بدست آمده استخراج گردید. بر روی پایگاه دادهٔ تلفنیِ FARSNUMBERS1 ، بهترین راندمانِ بازشناسیِ کلمات بازای داده‌های آزمایشی برای بازشناسی اعداد پیوستهٔ فارسی با واحدِ بازشناسیِ واج ۹۱/۱٪ بوده است. مقایسهٔ این راندمان با راندمان بازشناسیِ ارقامِ متصل با واحدِ بازشناسیِ کلمه (۸۳/۷٪)، بیانگر تأثیر چشمگیر انتخاب نوع واحدِ بازشناسی(واج یا کلمه) بر روی راندمان سیستم است.

قدردانی

این پروژه تحقیقاتی در راستای طرح ملی تحقیقات به شماره NRCI357 انجام و از طرف شورای پژوهشهای علمی کشور حمایت گردیده است.

مراجع

[۱] فرامرز فکری، محمدرضا نخعی، محمود تبیانی، شناسایی صحبت توسط کامپیوتر، دانشگاه صنعتی شریف، دانشکده مهندسی برق، پایان نامه کارشناسی ارشد، ۱۳۷۱.

[۲] حسن بابابیک، "بازشناسی گفتار با استفاده از تلفیق مدل مخفی مارکف و شبکه عصبی"، هفتمین کنفرانس مهندسی برق ایران، مرکز تحقیقات مخابرات ایران، تهران، ۱۳۷۸.

[۳] سعید بابایی‌زاده، ایمان غلامپور، کامبیز نایبی، "بهبود کارآیی سیستم‌های بازشناسی گفتار گسسته با ترکیب شبکه‌های عصبی و مدل‌های مارکف پنهان"، هفتمین کنفرانس مهندسی برق ایران، صص. ۱۸۳-۱۹۰، مرکز تحقیقات مخابرات ایران، تهران، ۱۳۷۸.

[۴] شیوا رستم‌زاده، سید محمد احدی، حمید شیخ‌زاده نجار، "بازشناسی گفتار فارسی ناپیوسته، به صورت ناوابسته به گوینده به کمک مدلهای پنهان مارکوف با چگالی پیوسته"، ششمین کنفرانس مهندسی برق ایران، صص. ۴-۹۳ تا ۴-۹۷، دانشگاه صنعتی خواجه نصیرالدین طوسی، تهران، ۱۳۷۷.

[۵] محمدمهدی همایون‌پور، امیر نجاری، "بازشناسی ارقام فارسی ناوابسته به گوینده با استفاده از مدل پیشگوی عصبی"، هفتمین کنفرانس مهندسی برق ایران، صص. ۷۵-۸۱ ، مرکز تحقیقات مخابرات ایران، تهران، ۱۳۷۸.

[۶] ابوالقاسم صیادیان، کامبیز بدیع، محمد حکاک، محمدرضا بیک‌زاده، "ارائه روش آماری FPG-GMM در بازشناسی گفتار"، هشتمین‌کنفرانس مهندسی برق ایران، صص. ۳۹۸-۴۰۶، دانشگاه صنعتی اصفهان، اصفهان، ۱۳۷۹.

[۷] احمد اکبری، بابک ناصرشریف، "بازشناسی هجاها در اعداد دورقمی فارسی بوسیله مدل مخفی مارکف"، ششمین کنفرانس سالانه انجمن کامپیوتر ایران، صص. ۴۳۲-۴۳۷، دانشگاه اصفهان، اصفهان، ۱۳۷۹.

[۸]علی طاهری دمنه، سید محمد احدی، سید علی سید صالحی، "بازشناسی گفتار پیوسته فارسی در دایره لغات متوسط به روش ترکیب شبکه های عصبی و مدلهای مارکف پنهان"، دهمین کنفرانس مهندسی برق ایران، دانشگاه تبریز، تبریز، ۱۳۸۱.

[8] A. K. Hunt, "New Commercial Applications of Telephone-Network-based Speech Recognition, and Speaker Verification", EUROSPEECH-91, pp. 431-433, Genova, Italy, 1991.

[9] J. Pencak, D. Nelson, "The NP Speech Activity Detection Algorithm", ICASSP-95, Vol. 1, pp. 381-384, May 1995.

[10] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction", IEEE Trans. ASSP, Vol. ASSP-27, No. 2, pp. 113-120, April 1979.

[11] R. J. Mammone, et al., "Robust Speaker Recognition: A Feature-based Approach", IEEE Signal Processing Magazine, pp. 58-71, Sep. 1996.

[12] L. R. Rabiner, B.-H. Juang, Fundamentals of Speech Recognition, Prentice-Hall, Englewood Cliffs, New Jersey, 1993.

[13] S. J. Young, P. C. Woodland, and W. J. Byrne, HTK: Hidden Markov Model Toolkit V1.5, Cambridge University, Engineering Department, Speech Group and Entropic Research Labs Inc., 1993.

[14] N. Morgan, H .Bourlard, "Neural Networks for Statistical Recognition of Continuous Speech", Proceedings of the IEEE, Vol. 83, No. 5, pp. 742-770, 1995.

[15] N. Morgan, H. Bourlard, "Continuous Speech Recognition: An Introduction to HMM/Connectionist Approach", IEEE Signal Processing Magazine, Vol. 12, pp. 25-42, 1995.

[16] N. Morgan, H. Bourlard, Connectionist Speech Recognition: A Hybrid Approach, Kluwer Academic Publishers, Boston, 1994.

[17] S. K. Riis, "Hidden Neural Networks: Application to Speech Recognition", ICASSP-98, Vol. 4, pp. 1117-1120, Seattle, USA, 1998.

[18] C. A. Ynoguti, E. da Silva Morais, F. Violaro, "A Comparison Between HMM and Hybrid ANN-HMM Based Systems for Continuous Speech Recognition", International Telecommunication Symposium, Vol. 1, pp. 135-140, Brazil, 1998.

[19] C. R. Jankowski, et al., "A Comparison of Signal Processing Front Ends for Automatic Word Recognition", IEEE Trans. on Speech and Audio Processing, Vol. 3, No. 4, pp. 286-293, July 1995.