

بازشناسی مقاوم گفتار با استفاده از ویژگی الگوهای زمانی به دست آمده از ساختار شبکه

عصبی بهینه شده MTMLP

یاسر شکفته^۱، فرشاد الماس گنج^۲

۱- مربی، گروه پردازش صوت و زبان طبیعی، پژوهشگاه توسعه فناوری‌های پیشرفته - تهران - ایران و

دانشجوی دکتری، دانشکده مهندسی پزشکی - دانشگاه صنعتی امیرکبیر - تهران - ایران

shekofteh@aut.ac.ir

۲- دانشیار گروه بیوالکتریک، دانشکده مهندسی پزشکی - دانشگاه صنعتی امیرکبیر - تهران - ایران

almas@aut.ac.ir

چکیده: ویژگی الگوهای زمانی سیگنال صوتی از دو حوزه زمانی و یا بردارهای بازنمایی شده قابل استخراج است. این ویژگی دربرگیرنده اطلاعات و مشخصات زمان بلند از تغییرات پیوسته واحدهای گفتاری است. در این مقاله، ویژگی الگوهای زمانی با استفاده از خروجی مقدار احتمال پسین واجی ساختار بهینه شده شبکه عصبی MTMLP، از مجموعه بردارهای بازنمایی مبتنی بر طیف (مانند ویژگی گفتاری LFBE) و همچنین، مبتنی بر کپستروم (مانند ویژگی گفتاری MFCC) استخراج شده است. با ترکیب اطلاعات الگوهای زمانی (دینامیک زمان بلند) به دست آمده از حوزه‌های لگاریتم طیف و کپستروم به بردار ویژگی‌های پایه بازشناسی، شامل ویژگی‌های گفتاری متداول MFCC و مشتقات زمانی اول و دوم آن (دینامیک زمان کوتاه)، نشان داده شده است که دقت بازشناسی واج در شرایط دادگان آزمون تمیز، حدود ۱ درصد نسبت به نتایج بهترین سیستم پایه بازشناسی بهبود می‌یابد. این در حالی است که ویژگی‌های به دست آمده از روش پیشنهادی، بازشناسی مقاومتری را در شرایط نویزی مختلف (تا حدود ۱۳ درصد) حاصل می‌نمایند که نشان دهنده مقاوم به نویز بودن روش پیشنهادی است.

واژه‌های کلیدی: بازشناسی گفتار، استخراج ویژگی، الگوهای زمانی، احتمال پسین، شبکه عصبی، مدل مخفی مارکوف.

۱- مقدمه

انتقال و ... نیز از دیگر حوزه‌های فعال در بحث بازشناسی گفتار است [۱]. بیشتر تحقیقات انجام شده در زمینه مقاوم‌سازی بازشناسی گفتار نسبت به تنوعات، روی سه تکنیک عمده بهسازی گفتار، استخراج ویژگی‌های مقاوم و جبران‌سازی پارامترهای مدل صوتی متمرکز شده است [۲]. از طرفی دیگر، تحقیقات اخیر نشان می‌دهد که نتایج به دست آمده از بهترین سیستم‌های ASR، پایین‌تر از نتایج بازشناسی سیستم شنوایی انسان است، از این رو، می‌توان امید داشت با الهام گرفتن از عملکرد فیزیولوژیک شنوایی انسان، بازشناسی این گونه سیستم‌ها را افزایش داد [۳]. برای نمونه، با در نظر گرفتن فرکانس مدولاسیون جهاز صوتی انسان در محدوده ۴ تا ۱۶ هرتز، محدوده زمانی مفید برای

در طی دو دهه اخیر محققان حوزه پردازش گفتار تلاش‌های زیادی برای بهبود عملکرد سیستم‌های خودکار بازشناس گفتار^۱ (ASR) در شرایط تمیز انجام داده‌اند. مقاوم‌سازی سیستم بازشناسی نسبت به تنوعات مختلف گفتاری (مانند تنوعات گوینده، لهجه، نویز محیط، کانال

^۱ تاریخ ارسال مقاله: ۱۳۹۲/۰۵/۱۹

تاریخ پذیرش مقاله: ۱۳۹۳/۰۲/۰۸

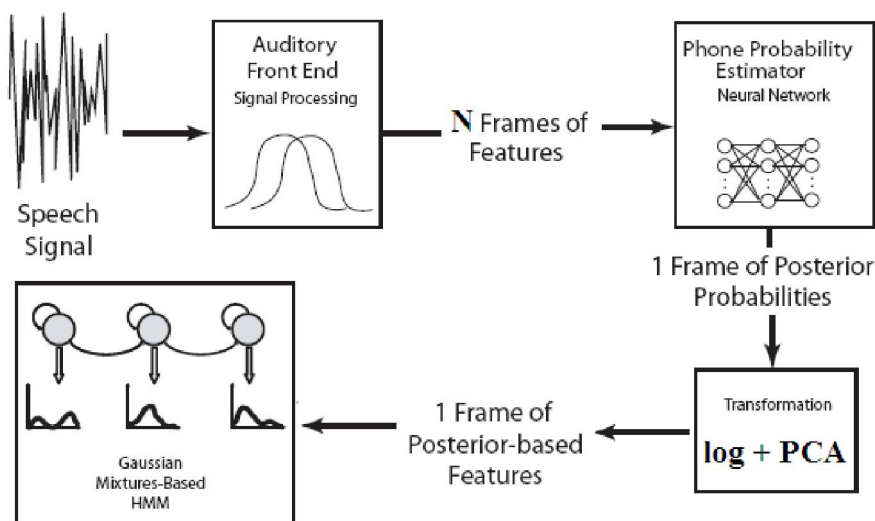
نام نویسنده مسؤول: یاسر شکفته

نشانی نویسنده مسؤول: ایران - تهران - خیابان حافظ - دانشگاه

صنعتی امیرکبیر - دانشکده مهندسی پزشکی

بازنمایی شده از سیگنال گفتار ارائه نمودند [۷]. آنها در این روش، به جای اعمال مستقیم بردارهای بازنمایی متداول گفتاری به سیستم بازشناس، با اعمال نگاشت غیرخطی شبکه عصبی بر روی توالی زمانی مربوط به هر یک از عناصر بردارهای بازنمایی شده و تخمین مقادیر احتمالات پسین^۳ (PP) واجی، در جهت تولید ویژگی‌های جدید اقدام نمودند. در ادامه و در [۸] دو مدل شبکه عصبی دیگر (HATS و TMLP)، در جهت تکمیل روش‌های استخراج ویژگی مبتنی بر الگوهای زمانی معرفی و بررسی شده است. ساختار سیستم بازشناس نهایی TANDEM نام داشت که نگاشت شبکه عصبی در بخش استخراج ویژگی آن و مدل مخفی مارکوف (HMM)، مدل بازشناس اصلی آن بود (شکل ۱). علاوه بر نگاشت غیرخطی شبکه عصبی، اثر استفاده از نگاشت‌های خطی آنالیز متمایزگر خطی^۴ (LDA) و آنالیز مؤلفه‌های اساسی^۵ (PCA) نیز بررسی شده است که نتایج ضعیف‌تری نسبت به نگاشت غیرخطی شبکه عصبی دربرداشتند [۹].

اطلاعات هر قاب گفتاری تا ۲۵۰ میلی ثانیه گسترش می‌یابد [۴]. از طرفی دیگر، با بررسی آلن (Alen) بر روی مدل چندباند درک آوا فلچر (Feltcher)، مشخص شد که استخراج اطلاعات صوتی در انسان، در باندهای مختلف فرکانسی انجام می‌گیرد. این پدیده با ساختار غشای قاعده‌ای درون حلزونی گوش، به عنوان یک آنالیزکننده طیفی قابل توجیه است [۵]. همچنین، ساختار سازمان‌های موازی و سلسله مراتبی در درک اطلاعات گفتاری انسان (واقع در بخش کورتکس شنوایی مغز) نیز یک‌سری پردازش‌های موازی و چندباند اطلاعات صوتی بین نرون‌های عصبی مختلف نشان می‌دهد. این نوع پردازش، به ترکیب مناسب اطلاعات کسب شده از هر باند فرکانسی منجر می‌شود که در نهایت به بازشناسی مقاوم‌تر انسان منجر خواهد شد [۶]. در سال ۱۹۹۹ میلادی، هرمانسکی (Hermansky) و شارما (Sharma) با الهام از شواهد فیزیولوژیک مطرح شده، یک روش جدید استخراج ویژگی مقاوم با عنوان TRAP (TempoRAI Pattern) به منظور استفاده از اطلاعات الگوهای زمانی^۲ (TP) موجود در توالی بردارهای



شکل (۱): ساختار سیستم بازشناس TANDEM [۶].

زیرباندها به صورت مستقل و یا ترکیبی انجام می‌شود، اما در روش TRAP، تعداد زیرباندها به تعداد عناصر بردار بازنمایی و البته، همراه با همپوشانی افزایش می‌یابد. همچنین، در روش TRAP از محدوده اطلاعات زمانی

از طرفی دیگر، ایده TRAP مشابه با روش چندباند است که در [۱۰] مطرح شده است. در روش چندباند، طیف فرکانسی مربوط به هر قاب زمانی به چند زیرباند بدون همپوشانی تقسیم و سپس استخراج ویژگی از هر یک از

آخر مقاله نتیجه‌گیری آورده شده است.

۲- استخراج ویژگی الگوهای زمانی

در روش استخراج ویژگی الگوهای زمانی، برخلاف سیستم‌های متداول بازشناسی گفتار که در آن ویژگی‌های استخراج شده براساس انرژی باندهای فرکانسی قاب‌های زمان کوتاه^۷ سیگنال گفتار به دست می‌آیند، اطلاعات مورد نیاز برای بازشناسی، از توالی هر یک از عناصر بردارهای بازنمایی (که ما آنها را دنباله عناصر ویژگی می‌نامیم) در یک محدوده زمانی نسبتاً طولانی‌تر حاصل می‌شوند. در شکل (۲) این تمایز نشان داده شده است که در آن روش TRAP یکی از روش‌های اولیه استخراج اطلاعات الگوی زمانی (TP) است [۷]. همان‌طور که از شکل (۲) استنباط می‌شود، ویژگی‌های به دست آمده از روش TRAP، بیانگر تغییرات دنباله هر یک از عناصر ویژگی خواهند بود. از این رو، این ویژگی، الگوی زمانی (TP) نامیده می‌شود. در حالت کلی این روش جزو روش‌های پس‌پردازش زمانی مبتنی بر داده^۸ محسوب می‌شود [۱۳، ۱۵].

مدلی که در ابتدا برای استخراج ویژگی TP پیشنهاد شده بود، مدل Neural TRAPs نام داشت که از دو طبقه شبکه عصبی چند لایه پرسپترون^۹ (MLP) تشکیل می‌شد [۷]. در طبقه اول این مدل، به تعداد عناصر (بعد) بردار بازنمایی گفتاری، شبکه عصبی MLP سه لایه به منظور یادگیری احتمال پسین واجی (خروجی شبکه) هر یک از دنباله‌های عناصر بردار ویژگی (ورودی شبکه) قرار داشت. در نتیجه، هر یک از MLP‌های تعلیم یافته در طبقه اول مدل، همانند یک فیلتر تطبیقی، اطلاعات TP مربوط به واج‌ها را از دنباله‌های مربوط به یک عنصر بردار ویژگی یاد می‌گرفت. در طبقه دوم مدل نیز با استفاده از یک شبکه عصبی دیگر، اطلاعات TP به دست آمده از خروجی MLP‌های طبقه اول، ترکیب می‌شد. بدین ترتیب، یک نگاهت با توانایی تخمین احتمال پسین مربوط به هر کلاس واجی از روی دنباله ویژگی‌های ورودی اعمالی به آن تولید می‌شد که مقید به یادگیری الگوی زمانی (TP) واج‌ها از دنباله‌های ویژگی ورودی به آن بود [۱۳].

بسیار بزرگتری نسبت به روش چند باند استفاده می‌شود، اما خاصیت مشترک هر دو روش در مقاوم بودن آنها در برابر تنوعات گفتاری است.

از جمله فعالیت‌های دیگری که به منظور استخراج و بهبود این نوع ویژگی انجام گرفته است، بهبود ورودی نگاهت غیرخطی شبکه عصبی است. در [۱۱] ورودی مورد نیاز برای مدل شبکه عصبی، با استفاده از اعمال مستقیم بانک فیلترهای میان‌گذر بر روی سیگنال زمانی گفتار انجام شده است. در [۱۲] نیز ورودی‌های شبکه با اعمال روش پیشگویی خطی (LP) بر طیف سیگنال تولید شده است. همچنین، هرمانسکی در [۱۳] نشان داد که استفاده از اطلاعات سه دنباله ویژگی مجاور هم به جای یک دنباله، باعث افزایش نتایج بازشناسی خواهد شد. روش‌هایی نیز برای بهبود ساختار شبکه عصبی مدل‌ها معرفی شده است. برای مثال، در [۱۴] اثر کاربرد شبکه‌های سلسله مراتبی^۶ بررسی شده است که به تعلیم مناسب‌تر واج‌های مشابه منجر می‌شود.

در تحقیق حاضر، روش بهبود یافته‌ای برای استخراج ویژگی‌های مقاوم گفتاری مبتنی بر ایده الگوهای زمانی پیشنهاد شده است. این روش شامل دو ایده پیشنهادی در تغییر ساختار لایه خروجی شبکه عصبی TMLP و ترکیب اطلاعات به دست آمده از خروجی دو شبکه است که این شبکه‌ها دربرگیرنده ویژگی‌های متداول از حوزه‌های متمایز کپستروم و طیفی هستند تا بتوانند در بهبود بخش نگاهت غیرخطی شبکه عصبی، به منظور افزایش کارایی سیستم بازشناسی گفتار با ساختار TANDEM مؤثر باشند. از این رو، در بخش ۲ به معرفی اولیه ویژگی الگوی زمانی (TP) و خواص ویژگی‌های گفتاری مبتنی بر احتمالات پسین خواهیم پرداخت. در بخش ۳ مجموعه دادگان و سیستم بازشناس معرفی می‌شود. بخش ۴ شامل ارائه مدل شبکه عصبی TMLP و مدل پیشنهادی MTMLP است. در بخش ۵ چگونگی اعمال تغییرات مورد نیاز بر روی ویژگی‌ها و نحوه اعمال آنها به سیستم بازشناس بیان می‌شود. در بخش ۶ نتایج آزمایش‌های به دست آمده از روش پیشنهادی ارائه و بحث و بررسی می‌شوند و در بخش

نامتغیر باشد [۲۶]. در پایان اینکه، خروجی‌های شبکه که شامل تخمینی از احتمالات پسین است، حاوی خصوصیات مفید (مانند مقدار مثبت و مجموع یک) است که یک چارچوب کارآمد برای ترکیب چندین کلاس‌بندی کننده ایجاد می‌کند [۲۷].

ویژگی‌های مبتنی بر مقادیر احتمال پسین با توجه به ذات احتمالاتی بودن آنها، کاربرد مناسبی در حوزه شناسایی الگو دارند. برخی از خواص مهم این ویژگی‌ها در ادامه آورده شده است:

الف) حساسیت کمتر به تغییرات غیرزبانی

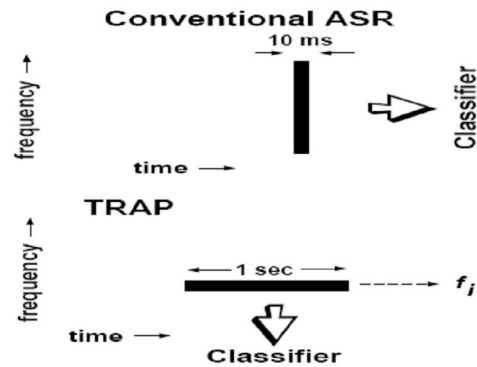
در ویژگی‌های متداول صوتی (مانند MFCC^{۱۱} یا LFBE^{۱۲} که مبتنی بر مدل منبع-فیلتر سیگنال گفتار هستند) درجه بالایی از تغییرات غیرزبانی^{۱۳} مانند مشخصه‌های گوینده و محیط (مانند نویز و کانال) دیده می‌شود. در [۲۸،۲۹] نشان داده شده است که اثرهای منفی مشخصه‌های هم‌تولیدی^{۱۴} در ویژگی‌های مبتنی بر احتمالات پسین نسبت به ویژگی‌های صوتی کمتر تأثیرگذار خواهد بود.

ب) خاصیت تُنکی

ویژگی‌های پسین حاوی احتمالات کلاس‌های واجی به شرط ویژگی‌های صوتی اعمال شده هستند. از این رو، برای هر قاب گفتاری، مجموع این احتمالات یک خواهد بود. به علاوه، این احتمالات به صورت تُنک^{۱۵} در فضای ویژگی پسین توزیع یافته‌اند. توزیع تُنک یکی از خواص مطلوب ویژگی‌های پسین است که در [۲۳] بررسی شده است. توضیحات مناسبی در مورد مقدار جرم چگالی احتمالاتی آنها و تمایز آنها در داده‌های تلفنی و میکروفونی در [۲۱] آورده شده است. شایان ذکر است که بیان ویژگی‌های یک سیگنال به صورت تُنک در کاربردهای فشرده‌سازی و مقاوم‌سازی آن نیز یکی از حوزه‌های مطالعاتی جذاب در دهه اخیر بوده است [۳۰،۳۳].

ج) تفکیک‌پذیری خطی بالا

این خاصیت در ساختار سلسله‌مراتبی که شامل چند طبقه متوالی از شبکه‌های عصبی است، بسیار مفید است. در این مورد، پارامترهای مدل شبکه عصبی در طبقات ثانویه ساختار سلسله‌مراتبی، باید به گونه‌ای بهینه‌سازی شوند که



شکل (۲): نحوه استفاده از ویژگی در روش‌های متداول بازشناسی گفتار (بالا) و روش TRAP (پایین) [۷].

به‌کارگیری ویژگی‌های مبتنی بر تخمین احتمال پسین یکی از حوزه‌های جدید در تحقیقات کاربردی برای بازشناسی گفتار است که در یک دهه اخیر مورد توجه محققان قرار گرفته است [۱۶،۲۲]. در این حوزه اغلب از کلاس‌بندی کننده غیرخطی مبتنی بر شبکه عصبی (مانند MLP) برای مدل‌سازی صوتی و تولید تخمین مقادیر احتمال پسین استفاده می‌شود [۲۳،۲۴]. در این مجموعه از روش‌ها، ورودی شبکه عصبی دربرگیرنده ویژگی‌های صوتی استاندارد همراه با محتوای زمانی اطراف هر قاب گفتاری است.

در [۲۵] نشان داده شده است که اگر یک مدل شبکه عصبی به خوبی بر روی حجم وسیع و متنوعی از دادگان تعلیم گفتاری آموزش یابد، می‌تواند در لایه خروجی خود، تخمین مناسبی از مقدار احتمال پسین کلاس‌های گفتاری واج یا حالت‌های واجی^{۱۶} را به شرط ویژگی‌های ورودی تولید نماید. مدل‌سازی صوتی مبتنی بر شبکه عصبی دارای مزیت‌هایی است: اول اینکه به فرض دقیق بر روی نحوه توزیع ویژگی‌ها و شکل پارامتری تابع چگالی آنها نیاز ندارد. در نتیجه، ویژگی‌های متنوع ورودی از کلاس‌های مختلف گفتاری که هرکدام دارای شکل توزیع متفاوتی هستند، می‌توانند به راحتی با یکدیگر ملحق و به عنوان ورودی شبکه استفاده شوند [۲۳]؛ دوم اینکه نشان داده شده است که اگر شبکه عصبی بر روی حجم وسیع و متنوعی از دادگان تعلیم آموزش یافته باشد، می‌تواند نسبت به مشخصه‌های گوینده و اطلاعات خاص محیطی مانند نویز

بردار بازنمایی ۱۹ عنصری LFBE، از اعمال تابع لگاریتم، بر روی انرژی بانک فیلترهای ۱۸ تایی به دست آمده در مقیاس غیرخطی مل (Mel) به همراه ویژگی انرژی کل طیف (E0) استفاده شده است. بردار ۱۳ عنصری MFCC نیز با استفاده از ۱۲ ضریب اول کپستروم حاصل از ویژگی‌های LFBE و همچنین، ضریب صفرم کپسترال (C0) به دست آمده است. در بخش استخراج بردارهای بازنمایی سیگنال‌های گفتاری، از قاب‌های گفتاری با طول زمانی ۲۳/۲ میلی ثانیه (۵۱۲ نمونه از سیگنال در هر قاب گفتاری) و همپوشانی ۵۰٪ استفاده شده است. مقدار ضریب پیش‌تاکید نیز برابر ۰/۹۷۵ در نظر گرفته شد. پس از تولید بردارهای بازنمایی، روش تفریق میانگین (MS) در جهت مقاوم‌سازی بیشتر ویژگی‌ها اعمال شده است [۳۷]. از طرفی دیگر، به منظور مقایسه عملکرد ویژگی‌های به دست آمده از الگوهای زمانی (که بیانگر دینامیک زمان‌بلند دنباله‌های ویژگی هستند) با ویژگی‌های دینامیک مشتقات اول و دوم بردار ویژگی‌ها (که بیانگر دینامیک زمان‌کوتاه بردار ویژگی‌ها هستند)، مشتقات اول و دوم بردارهای بازنمایی نیز در این مرحله محاسبه می‌شوند.

از ساختار سیستم بازشناس Tandem معرفی شده در بخش ۱ همراه با مدل مخفی مارکوف (HMM) (به عنوان مدل بازشناس واج به وسیله نرم افزار HTK [۳۸]) برای ارزیابی کارایی روش استخراج ویژگی پیشنهادی استفاده شده است. در این جهت برای هر واج، یک مدل از چپ به راست با سه حالت و هر حالت شامل مخلوط ۱۶ مدل گوسی (GMM) در نظر گرفته شده است. نتایج بازشناسی ارائه شده به صورت درصد دقت بازشناسی واج (Acc%) و به صورت بازشناسی پیوسته و مستقل از گوینده خواهند بود.

دادگان گفتاری مورد استفاده، از مجموعه دادگان فارسی‌دات میکروفونی کوچک با نرخ نمونه‌برداری ۲۲۰۵۰ هرتز است [۳۹]. از آنجایی که فایل‌های صوتی دادگان

خطای بین بردارهای احتمالاتی پسین تخمین زده شده (خروجی شبکه طبقه اول به عنوان بردار ویژگی برای شبکه طبقه دوم) و بردارهای هدف خروجی (که به طور متداول در شکل صفر و یک یا همان قالب One-Hot هستند) کمینه شود. بردارهای هدف کلاس‌های واجی، در فضای چندبُعدی ویژگی‌های پسین، تفکیک‌پذیری خطی آنها را بیشتر مهیا می‌سازد [۲۱]. اگر الگوریتم تعلیم شبکه براساس کمینه‌سازی میانگین مجذور خطا^{۱۶} (MSE) باشد، تخمینی از مقدار احتمالاتی پسین کلاس‌های واجی در خروجی شبکه به شرط قطعه^{۱۷} قاب‌های گفتاری ورودی اعمال شده به آن تولید خواهد شد [۳۴،۳۵].

۳- معرفی روش‌های متداول بازنمایی،

دادگان و سیستم بازشناس مورد استفاده

در این مقاله، استخراج ویژگی‌های احتمالات پسین مبتنی بر الگوهای زمانی، علاوه بر اینکه از مجموعه بردارهای بازنمایی لگاریتم انرژی فیلتر بانک (ویژگی‌های LFBE که در حوزه طیف هستند) انجام شده است، از مجموعه ضرایب کپسترال (ویژگی‌های MFCC که در حوزه کپستروم قرار دارند) نیز محاسبه شده است. در [۳۶] نشان داده شده است که با متوسط‌گیری مقادیر احتمالاتی پسین به دست آمده از دو مدل شبکه عصبی (خروجی هر شبکه) که یکی بر روی ویژگی‌های حوزه لگاریتم طیف و دیگری بر روی ویژگی‌های حوزه کپستروم تعلیم یافته‌اند، بهبود نتایج خروجی شبکه عصبی حاصل خواهد شد؛ خصوصاً که ویژگی‌های حوزه لگاریتم طیف در شرایط تمیز و کم‌نویز و ویژگی‌های کپستروم در شرایط نویزی‌تر عملکرد بهتری دارند. بنابراین، با این شیوه ترکیب، می‌توانیم به دستیابی یک تخمین احتمال پسین مقاوم‌تر در شرایط مختلف تمیز و نویزی امیدوار باشیم.

از این‌رو، بردارهای بازنمایی مورد استفاده، شامل ضرایب MFCC و LFBE هستند که برای به دست آوردن

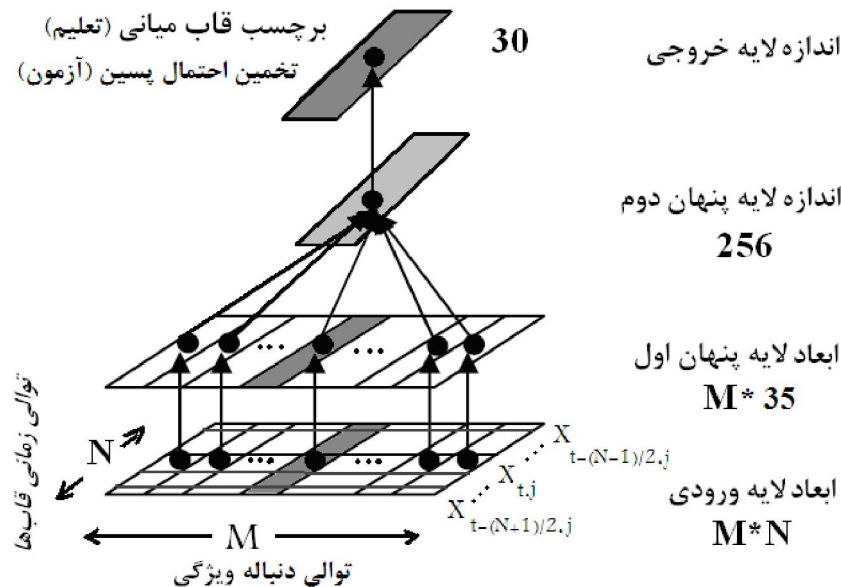
۴- روش استخراج ویژگی TP به وسیله مدل

TMLP از بردارهای بازنمایی

یکی از روش‌های مناسب برای استخراج ویژگی الگوهای زمانی، استفاده از مدل شبکه عصبی است. در این تحقیق از مدل TMLP^{۱۸} که ساختار آن الهام گرفته شده از بخش تونوتوپیک^{۱۹} سیستم شنوایی انسان است، در جهت استخراج ویژگی الگوهای زمانی استفاده می‌شود [۸]. این مدل که در شکل (۳) نشان داده شده است، برخلاف مدل Neural TRAPS، تنها شامل یک طبقه MLP چهار لایه است و بنابراین، آموزش آن تنها با یک مرحله تعلیم انجام می‌گیرد. این نوع ساختار مدل شبکه عصبی باعث می‌شود که در لایه‌های پایینی شبکه، پردازش ویژگی‌های ورودی به طور مستقل انجام گرفته و سپس این اطلاعات در لایه‌های بالاتر شبکه ترکیب شوند.

مورد استفاده در شرایط اتاق سکوت ضبط شده‌اند، دارای نرخ سیگنال به نویز (SNR) حدود 34dB هستند. بنابراین برای تولید دادگان نویزی با نویز جمع شونده، از نویزهای همهمه (نویز واقعی پیش زمینه)، صورتی (نویز باند باریک) و سفید (نویز باند پهن) مجموعه دادگان نویزی "NOISEX-92" در مقادیر مختلف سیگنال به نویز استفاده می‌شود [۴۰].

مجموعه دادگان مورد استفاده شامل ۶۰۶۰ جمله است که از ۵۰۰۰ جمله (حدود ۴ ساعت) به عنوان دادگان تعلیم و از مابقی آن (حدود یک ساعت) برای دادگان آزمون استفاده شده است. همچنین، به منظور تعمیم‌پذیری بیشتر مدل بازشناس نهایی، از ۲۰۰۰ جمله اول مجموعه دادگان تعلیم برای آموزش مدل شبکه عصبی و از ۳۰۰۰ جمله دیگر برای تعلیم مدل مخفی مارکوف استفاده شده است.



شکل (۳): ساختار مدل TMLP با توالی N قاب زمانی و بردار ویژگی ورودی M بُعدی برای هر قاب زمانی.

برگیرنده اطلاعات مفیدتری از الگوهای زمانی باشند. از این رو، مدل TMLP مقید به یادگیری الگوهای زمانی (TP) موجود در هر یک از دنباله‌های ویژگی (چنانکه در تعلیم مدل Neural TRAPS اتفاق می‌افتد) نیست. ساختار مدل

تفاوتی که میان عملکرد این دو نوع مدل وجود دارد این است که در مدل TMLP به علت پسانتشار خطای ناشی از الگوریتم تعلیم بر روی تمامی دنباله‌های ویژگی ورودی آن، این مدل دنباله‌های ویژگی را یاد می‌گیرد که در

بعد قاب میانی نیز در لایه خروجی شبکه استفاده کنیم. با این روش، تعداد نرون‌های لایه خروجی به ۹۰ نرون (سه خروجی که هر کدام شامل ۳۰ نرون است) افزایش می‌یابد که در هنگام محاسبه مقادیر احتمالات پسین، می‌توان میانگین وزن‌داری از احتمالات مربوط به قاب‌های قبل و بعد را به قاب میانی افزود. این ساختار پیشنهادی را که می‌تواند به هموارسازی نتایج احتمالاتی خروجی کمک نماید (و در نتیجه تولید ویژگی‌هایی که می‌تواند به کاهش اثر درج واج منجر شوند)، مدل بهبودیافته TMLP (MTMLP) می‌نامیم. برای بیان کمی بهبود نتایج خروجی مدل MTMLP نسبت به مدل TMLP (که معادل با تمایزپذیری بیشتر ویژگی الگوی زمانی به دست آمده است)، از معیار دقت بازشناسی قاب استفاده می‌نماییم. در این معیار، با توجه به مقادیر احتمالاتی پسین به دست آمده برای هر قاب، کلاس واجی که بیشترین احتمال را کسب نموده، به عنوان برچسب قاب تعیین می‌شود. سپس این برچسب با برچسب واقعی قاب مقایسه می‌شود. درصد دقت بازشناسی قاب، از نسبت تعداد برچسب‌های درست تخمین زده شده به مجموع تعداد تمامی قاب‌ها محاسبه خواهد شد. در جدول (۱) نتایج دقت بازشناسی قاب مدل‌های TMLP و MTMLP با استفاده از بردارهای بازنمایی ورودی MFCC بر روی مجموعه دادگان آزمون تمیز آورده شده است.

جدول (۱): درصد دقت بازشناسی قاب با استفاده از مدل‌های TMLP و MTMLP بوسیله ویژگی‌های MFCC.

مدل	کل واجها	واکه	شبه واکه	انفجاری	سایشی
TMPLP	۷۱/۲۶	۷۵/۸۷	۴۷/۲۲	۶۹/۰۳	۴۵/۳۸
MTMLP	۷۲/۲۷	۷۸/۰۲	۵۵/۵۶	۴۷/۷۵	۶۹/۴۲

آن‌گونه که از نتایج جدول ۱ پیداست، با پیاده‌سازی مدل MTMLP، دقت بازشناسی قاب ۷۲/۲۷ درصد بر

TMPLP، برای ۱۳ ویژگی MFCC به صورت $256 \times 30 - 13 \times (35 \times 256) - 13 \times (21 \times 35)$ در نظر گرفته شده است. بنابراین، محدوده زمانی بردارهای ویژگی ورودی به شبکه برای استخراج ویژگی الگوی زمانی در حدود ۲۵۰ میلی‌ثانیه $(11.6 \times (21+1))$ خواهد بود.

تعلیم شبکه براساس برچسب‌دهی باینری نوع سخت (One Hot) انجام شده است. در این نوع برچسب‌دهی، به ازای هر مجموعه بردار بازنمایی ورودی به شبکه، یک خروجی ۳۰ نرونی (به تعداد کلاس‌های واجی) به عنوان خروجی مطلوب آن تعریف می‌شود که یک نرون آن مقدار یک (متناظر با شماره کلاس واج مربوط به قاب میانی مجموعه بردار ورودی) و بقیه نرون‌های آن مقدار صفر دارند. بنابراین، پس از مرحله تعلیم، لایه خروجی شبکه می‌تواند بیانگر تخمین مقدار احتمالاتی پسین $P(\omega_i | \bar{x}_j)$ هر یک از ۳۰ کلاس واج فارسی $(\omega_i, i = 1, \dots, 30)$ ، برای بردار ویژگی میانی \bar{x}_j ، از مجموعه بردارهای $[\bar{x}_{j-(N-1)/2}, \dots, \bar{x}_{j+(N-1)/2}]$ اعمالی به ورودی شبکه باشد. در ادامه، با اعمال یک‌سری تبدیلات خطی و غیرخطی بر روی مقادیر احتمالاتی پسین به دست آمده، بردار ویژگی جدید (که دربرگیرنده اطلاعات الگوهای زمانی واج‌هاست) تشکیل می‌شود. از این‌روست که این روش، روش استخراج ویژگی با استفاده از پردازش زمانی مبتنی بر داده نامیده می‌شود.

۴-۱- مدل پیشنهادی MTMLP و بهبود روش استخراج ویژگی TP

با توجه به دیدگاه الگوهای زمانی، از آن‌جا که اطلاعات موجود در ورودی مدل شبکه عصبی TMLP، به تعداد زیادی از بردارهای بازنمایی مربوط هستند (در اینجا ۲۱ بردار) بنابراین، برای تعلیم مناسب‌تر نگاشت شبکه عصبی پیشنهاد می‌شود که علاوه بر استفاده از برچسب واج قاب میانی ورودی، از اطلاعات واجی مربوط به قاب‌های قبل و

جدول (۲): درصد دقت بازشناسی قاب مدل MTMLP برای ۱۳ ویژگی MFCC و ۱۹ ویژگی LFBE و همچنین مدل ترکیبی آن دو.

سایشی	انفجاری	شبه واکه	واکه	کل واجها	ویژگی مدل
۶۹/۴۲	۴۷/۷۵	۵۵/۵۶	۷۸/۰۲	۷۲/۲۷	MFCC13
۷۲/۷۴	۵۳/۹۰	۶۳/۲۳	۷۸/۴۱	۷۴/۱۷	LFBE19
۷۴/۶۶	۵۶/۰۹	۶۳/۸۰	۸۱/۱۳	۷۶/۳۸	ترکیبی

در جدول (۲)، نتایج بازشناسی قاب مدل MTMLP با استفاده از ویژگی های کپستروم (MFCC13) و مدل ترکیبی آن دو نیز آورده شده است. مدل ترکیبی مورد استفاده (ترکیب در سطح خروجی)، متشکل از دو مدل تعلیم یافته MTMLP بر روی ویژگی های LFBE و MFCC است که در آن مقدار احتمال پسین برای هر قاب ورودی $P(\omega_i | \bar{x}_j)$ از متوسط گیری مقادیر احتمالاتی نرمالیزه به دست آمده از لایه خروجی هر یک از شبکه های فوق حاصل شده است که در رابطه (۱) آورده شده است.

$$P(\omega_i | \bar{x}_j) = 0.5\{P_1(\omega_i | \bar{x}_{MFCC,j}) + P_2(\omega_i | \bar{x}_{LFBE,j})\} \quad (1)$$

با توجه به نتایج جدول ۲، مدل بازشناس ترکیبی توانسته است برای تمامی کلاس های واجی، به بهبود دقت بازشناسی قاب (تشخیص دقیق تر احتمالات پسین) منجر شود. بهبود نتایج مدل ترکیبی نسبت به مدل با ویژگی های ورودی LFBE حدود ۲/۲ درصد و نسبت به مدل با ویژگی های ورودی MFCC حدود ۴/۱ درصد است.

۵- آماده سازی ویژگی های استخراج شده

پیش از استفاده از مقادیر احتمالاتی پسین به دست آمده به عنوان ویژگی الگوی زمانی (TP) در ساختار TANDEM با مدل بازشناس HMM، باید یک سری تبدیلات مفید بر روی آنها اعمال شود. در ابتدا برای توزیع مناسبتر این مقادیر، مقدار میانگین هر بردار را صفر

روی دادگان آزمون تمیز حاصل می شود که بازشناسی دقیقتری (حدود یک درصد) نسبت به مدل اولیه TMLP دربرداشته است. همچنین، مدل MTMLP به تخمین بهتر احتمالات پسین واج های با طول زمانی بلند (واکه ها، شبه واکه ها و سایشی ها) منجر شده است.

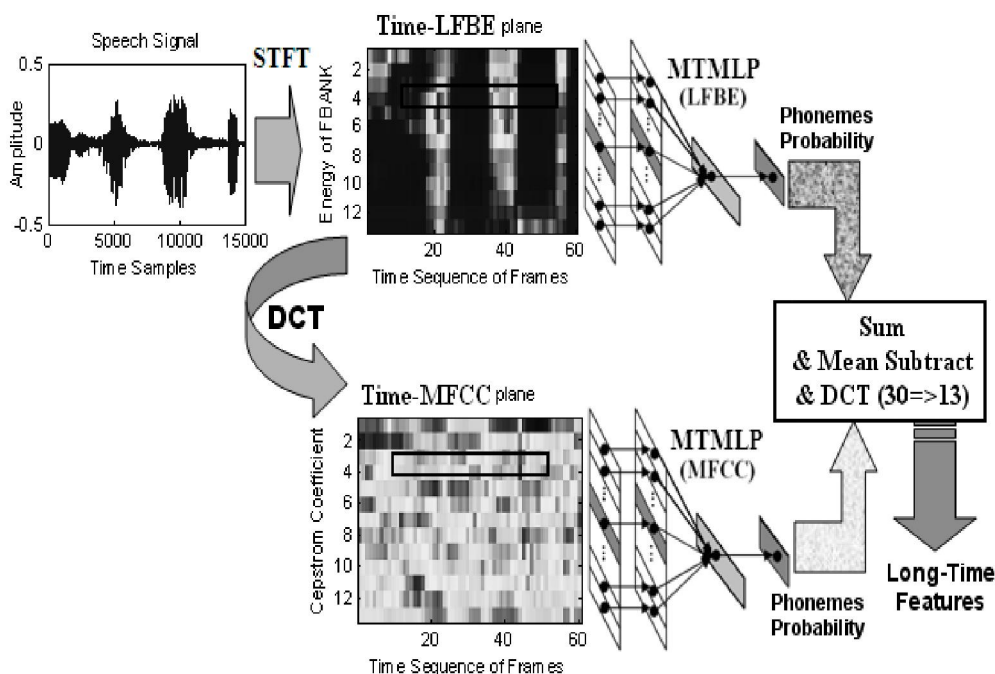
۴-۲- ترکیب اطلاعات به دست آمده از

بردارهای بازنمایی طیفی و کپستروم

استفاده همزمان از بردارهای ویژگی متمایز که حاوی اطلاعات متفاوتی از یک سیگنال هستند، می تواند به افزایش کارایی عملکرد سیستم های بازشناسی منجر شود [41-43]. در این بخش نشان خواهیم داد که چگونه ترکیب نتایج خروجی به دست آمده از شبکه های تعلیم یافته با بردارهای بازنمایی مبتنی بر طیف و کپستروم، به بهبود نتایج دقت بازشناسی قاب مدل بازشناس شبکه عصبی منجر می شود. برای این منظور مدل پیشنهادی MTMLP را علاوه بر بردارهای بازنمایی حوزه طیف (LFBE) تعلیم می دهیم. البته، ساختار مدل شبکه عصبی برای ویژگی های LFBE به گونه ای انتخاب می شود که به تعداد پارامترهای مدل تعلیم یافته با ویژگی MFCC، پارامتر وزن وجود داشته باشد. از این رو، ساختار شبکه عصبی پیشنهادی این مدل برای ویژگی های حوزه طیف LFBE به صورت $19*(21*24)-19*(24*256)-256*90$ طراحی شد. به این ترتیب، برای هر دو مدل شبکه عصبی، نسبت تعداد کل قاب های تعلیمی به تعداد کل وزن های مجهول شبکه، یکسان و برابر مقدار $6/3$ خواهد بود (برای مناسب بودن خاصیت تعمیم پذیری و همچنین زمان آموزش مناسب شبکه، مقدار این نسبت باید بین ۴ تا ۱۰ باشد [۴۴]). در جدول (۲) نتایج دقت بازشناسی قاب مدل MTMLP با ویژگی های طیفی (LFBE19) آورده شده است.

استفاده در هر حالت مدل HMM منجر می‌شود) استفاده می‌نماییم [۲۲]. در شکل (۴) نحوه استخراج ویژگی الگوهای زمانی در روش پیشنهادی نشان داده شده است.

می‌نماییم. در مرحله بعد از تابع لگاریتم برای مقیاس کردن غیرخطی و تبدیل کسینوسی گسسته^{۲۰} (DCT) در جهت کاهش بُعد و غیرهمبسته نمودن آنها (که به همگونی بیشتر توزیع ویژگی‌های به دست آمده، با مدل‌های گوسی مورد



شکل (۴): روش پیشنهادی برای استخراج ویژگی TP.

۶- پیاده‌سازی آزمایش‌ها و بحث و بررسی

۶- پیاده‌سازی آزمایش‌ها و بحث و بررسی

در این بخش، با تعریف مجموعه بردار ویژگی‌های متمایز که هر یک به صورتی متمایز از ویژگی الگوهای زمانی (TP) به دست آمده استفاده می‌نمایند، سیستم‌های بازشناس HMM را به‌طور جداگانه تعلیم داده، سپس برای مقایسه میزان کارایی آنها، نتایج بازشناسی هر یک از بردارهای ویژگی را ارائه می‌نماییم. در جدول (۳) بردار ویژگی‌های تشکیل شده برای هر سیستم شرح داده شده است. همچنین، در این جدول دو سیستم پایه HMM، حاوی ویژگی‌های متداول گفتاری MFCC و LFBE نیز تعریف شده‌اند که می‌توانند معیار مناسبی برای مقایسه نتایج بازشناسی با ویژگی‌های جدید معرفی شده باشند.

جدول (۳): تعریف سیستم‌های بازشناس به همراه خصوصیات

نام سیستم	خصوصیات بردار ویژگی
M1	۱۳ ویژگی استاتیک MFCC + مشتقات زمانی اول و دوم بردارهای ویژگی
M2	۱۹ ویژگی استاتیک LFBE + مشتقات زمانی اول و دوم بردارهای ویژگی
M3	۱۳ ویژگی استاتیک MFCC + مشتقات زمانی اول و دوم بردارهای ویژگی + ۱۳ ویژگی الگوهای زمانی
M4	۱۹ ویژگی استاتیک LFBE + مشتقات زمانی اول و دوم بردارهای ویژگی + ۱۳ ویژگی الگوهای زمانی
M5	۱۳ ویژگی استاتیک MFCC + مشتقات زمانی اول و دوم بردارهای ویژگی + ۱۳ ویژگی الگوهای زمانی MFCC + ۱۳ ویژگی الگوهای زمانی LFBE
M6	۱۳ ویژگی استاتیک MFCC + مشتقات زمانی اول و دوم بردارهای ویژگی + ۱۳ ویژگی الگوهای زمانی مدل های ترکیبی با ویژگی های MFCC و LFBE

استفاده از ویژگی الگوی زمانی (TP) که حاوی اطلاعات دینامیک زمان‌بلند هر دنباله ویژگی است، به علت دربرداشتن اطلاعات متمایزکننده، بهبود نتایج بازشناسی را در پی داشته است. این بهبود درحالی به دست آمده که از ویژگی‌های دل‌تا و دلتادلتا بازنمایی (دینامیک زمان‌کوتاه بردار ویژگی‌ها) نیز در سیستم‌های مورد نظر استفاده شده است.

از طرفی دیگر، با توجه به نتایج به دست آمده در جداول (۲) و (۴)، اگرچه ویژگی‌های LFBE نسبت به ویژگی‌های MFCC بازشناسی قاب‌بیشتری با استفاده از مدل شبکه عصبی MTMLP داشتند، اما نتایج دقت بازشناسی واج کمتری با سیستم بازشناس HMM در پی خواهند داشت. ویژگی‌های MFCC با ویژگی‌های LFBE تنها در یک تبدیل DCT تفاوت دارند، بنابراین، استفاده از تبدیل DCT بر روی ویژگی‌های LFBE و تبدیل آنها به ویژگی‌های MFCC به بهبود بازشناسی منجر شده است، زیرا این تبدیل توانسته است ویژگی‌های اعمالی به مدل HMM را غیرهمبسته‌تر و در نتیجه تطابق بیشتری با فرض قطری بودن ماتریس کوواریانس مورد استفاده در هر عنصر گوسین مدل GMM در حالت‌های واجی HMM داشته باشد.

با مقایسه نتایج بازشناسی دو سیستم بازشناس M5 و M6 که به گونه‌ای متفاوت از اطلاعات الگوهای زمانی استفاده کرده‌اند، این نتیجه حاصل می‌شود که بیشتر از آن‌که الگوهای زمانی هر یک از بردار ویژگی‌های MFCC یا LFBE مفید باشند (سیستم M5)، ترکیب نتایج احتمالاتی پسین آن دو (سیستم M6) به بهبود بازشناسی منجر خواهد شد. با بررسی جداگانه نتایج درصد دقت بازشناسی قاب برای هر واج با مدل شبکه عصبی MTMLP، دیده شد که برخی از واج‌ها با استفاده از ویژگی‌های MFCC (مانند واج‌های انفجاری /ب/ و /ت/ و انفجاری-سایشی /ج/ و مدل سکوت) و برخی دیگر با استفاده از ویژگی‌های LFBE (مانند واج‌های شبه‌واکه /ی/، /ر/، /ل/، /م/ و سایشی /ف/، /ز/ و /ژ/) بهتر تشخیص داده می‌شوند، اما با استفاده از مدل ترکیبی آن دو، درصد دقت بازشناسی قاب اغلب واج‌ها (خصوصاً واکه‌ها و سایشی‌ها) افزایش می‌یابد. از این رو، سیستم پیشنهادی M6 می‌تواند با داشتن اطلاعات

در جدول (۴) نیز نتایج درصد دقت بازشناسی واج از آزمون بازشناسی واج پیوسته متناظر با سیستم‌های بازشناس تعریف شده در جدول (۳) آورده شده‌اند.

جدول (۴): درصد دقت بازشناسی واج سیستم‌های جدول (۳)

طول بردار ویژگی	۳۹	۵۷	۵۲	۷۰	۶۵	۵۲
سیستم	M1	M2	M3	M4	M5	M6
تمیز	۶۸/۷	۶۲/۲	۶۶/۸	۶۷/۳	۶۷/۸	۶۹/۶
نویز هممه						
SNR=20dB	۶۰/۹	۵۰/۶	۵۸/۳	۵۹/۲	۶۰/۴	۶۲/۶
SNR=10dB	۴۳/۴	۲۵/۷	۴۲/۶	۴۰/۱	۴۴/۱	۴۷/۳
SNR=0dB	۱۹/۴	۱۰/۰	۱۶/۴	۱۲/۳	۱۳/۶	۱۹/۳
نویز صورتی						
SNR=20dB	۶۳/۵	۵۶/۰	۶۳/۰	۶۳/۱	۶۴/۴	۶۶/۳
SNR=10dB	۴۷/۶	۳۵/۲	۵۳/۲	۴۸/۹	۵۳/۸	۵۶/۳
SNR=0dB	۲۵/۱	۲۱/۵	۳۶/۶	۳۰/۷	۳۳/۱	۳۸/۳
نویز سفید						
SNR=20dB	۶۴/۴	۵۶/۳	۶۳/۶	۶۴/۱	۶۵/۶	۶۷/۲
SNR=10dB	۴۹/۷	۳۸/۶	۵۴/۲	۵۱/۰	۵۵/۴	۵۷/۸
SNR=0dB	۳۰/۲	۲۵/۶	۳۸/۰	۳۳/۹	۳۶/۶	۴۱/۰

با توجه به نتایج جدول ۴ سیستم M6 (روش پیشنهادی نهایی) برای اغلب شرایط تمیز و نویزی نسبت به دیگر سیستم‌ها، دقت بازشناسی واج بالاتری کسب کرده است. برای مثال، در شرایط تمیز، این سیستم به بهبود حدود یک درصدی نسبت به بردار ویژگی پایه مبتنی بر MFCC (سیستم M1) و همچنین بهبود ۷/۴ درصدی نسبت به بردار ویژگی مبتنی بر LFBE (سیستم M2) منجر شده است. در شرایط نویز شدید 0dB هم این روش نسبت به نتایج بردار ویژگی‌های MFCC (که بازشناسی نویزی بهتری نسبت به ویژگی‌های LFBE دارند) برای نویز صورتی بهبود ۱۳/۲ درصدی و در نویز سفید بهبود ۱۰/۸ درصدی به دست آورده است، اما در این شرایط (نویزی شدید)، بهبودی برای نویز هممه (که یکی از سخت‌ترین شرایط نویزی در حوزه بازشناسی گفتار است) مشاهده نشده است. در مجموع،

تبدیلات مناسب بر روی مقادیر احتمالات پسین خروجی از مدل بهینه شده شبکه عصبی، ویژگی‌های به دست آمده را برای اعمال به سیستم بازشناس HMM (سیستم پیشنهادی M6) مهیا نمودیم. همچنین، نشان داده شد که با استفاده از ویژگی‌های جدید، نتایج بازشناسی بالاتری در اغلب شرایط تمیز و نویزی آزمون حاصل خواهند شد که نشان دهنده مقاوم به نویز بودن این روش است. این در حالی است که هزینه محاسباتی مربوط به روش استخراج ویژگی پیشنهادی ۲/۵ برابر روش استخراج ویژگی متداول MFCC است.

مراجع

- [1] Lippmann, R., "Speech Perception by Humans and Machines", Speech Communication, Vol. 22, No. 1, pp. 1-15, 1997.
- [2] Chulhee, L., "Optimizing Feature Extraction for Speech Recognition", IEEE Trans. on Speech and Audio Processing, Vol. 11, No. 1, pp. 80-87, 2003.
- [3] Deng, L., "Processing of Acoustic Signals in a Cochlear Model Incorporating Laterally Coupled Suppressive Elements", Journal of Neural Networks, Vol. 5, No. 1, pp. 19-34, 1992.
- [4] Drullman, R., Festen, J., Plomp, R., "Effect of Temporal Envelope Smearing on Speech Reception", Journal of the Acoustical Society of America, Vol. 95, No. 2, pp. 2670-2680, 1994.
- [5] Allen, J.B., "Harvey Fletcher's Role in the Creation of Communication Acoustics", Journal of the Acoustical Society of America, Vol. 99, No. 4, pp. 1825-1839, 1996.
- [6] Kandel, E., Essential of Neural System, Addison-Wesley Publishing Company, 1st Edition, 2002.
- [7] Hermansky, H., Sharma, S., "Temporal Patterns (TRAPS) in ASR of Noisy Speech", In Proc. ICASSP, Arizona, USA, pp. 289-292, 1999.
- [8] Chen, B., Zhu, Q., Morgan, N., "Tonotopic Multi-Layer Perceptron, a Neural Network for Learning Long-term Temporal Features for Speech Recognition", In Proc. ICASSP, USA, pp. 945-948, 2005.
- [9] Chen, B., Zhu, Q., Morgan, N., "Learning long-term Temporal Features in LVCSR using Neural Networks", In Proc. ICSLP,

متمايزتر (نشأت گرفته از ویژگی الگوی زمانی)، نسبت به دیگر سیستم‌ها درصد دقت بازشناسی واج بالاتری کسب نماید.

برای بررسی هزینه محاسباتی و سرعت انجام پیاده‌سازی برای تولید بردار ویژگی از فاکتور زمان حقیقی^{۲۱} در بخش استخراج ویژگی (FE-RTF) استفاده شده است [۲۵]. این فاکتور بیانگر نسبت زمان لازم برای پردازش یک فایل صوتی و تولید بردارهای ویژگی از آن به مدت زمان آن فایل صوتی است. در یک آزمون بر روی سیستم PC با مشخصات پردازنده 3GHz و حجم حافظه RAM برابر با 2G و با شرایط یکسان، سیستم پایه M1 مقدار FE-RTF=0.04 و روش پیشنهادی با سیستم M6 مقدار FE-RTF=0.10 به دست آمده است. بنابراین، هزینه محاسباتی در تولید بردار ویژگی در روش پیشنهادی حدود ۲/۵ برابر روش پایه و متداول MFCC است.

۷- نتیجه گیری

در این مقاله به معرفی و بررسی ویژگی الگوی زمانی (TP) و ویژگی‌های مبتنی بر تخمین احتمالات پسین در کاربردهای بازشناسی گفتار پرداختیم. نشان داده شد که ویژگی الگوی زمانی بیانگر دینامیک زمان‌بلند هر دنباله ویژگی است و علاوه بر اثرهای مفید استفاده از ویژگی مشتقات زمانی بردارهای بازنمایی که بیانگر دینامیک زمان کوتاه بردارهای ویژگی است، بهبود بیشتری را در نتیجه بازشناسی به دست خواهند آورد.

در این جهت، ابتدا با پیشنهاد مدل بهبودیافته شبکه عصبی MTMLP، نتایج تشخیص احتمالات واجی مدل شبکه عصبی TMLP را بهبود دادیم. سپس با استفاده از تعریف مدلی ترکیبی، اطلاعات الگوهای زمانی به دست آمده از مجموعه بردارهای ویژگی MFCC و LFBE را ترکیب نمودیم و نشان دادیم با استفاده از این روش نیز تشخیص احتمالات واجی بهبود می‌یابد. در ادامه، با اعمال

- Magimai, M., "Phase autocorrelation (PAC) Features for Noise Robust Speech Recognition", *Speech Communication*, Vol. 54, No. 7, pp. 867-880, 2012.
- [23] Zhu, Q., Chen, B., Morgan, N., Stolcke, A., "On Using MLP Features in LVCSR", In *Proc. ICSLP*, 2004.
- [24] Morgan, N., Chen, B.Y., Zhu, Q., Stolcke, A., "TRAPPING Conversational Speech: Extending TRAP/Tandem Approaches to Conversational Telephone Speech Recognition", In *Proc. ICASSP*, pp. 536-539, 2004.
- [25] Richard, M.D., Lippmann, R.P., "Neural Network Classifiers Estimate Bayesian a Posteriori Probabilities", *Neural computation*, Vol. 3, No. 4, pp. 461-483, 1991.
- [26] Ikbali, S., *Nonlinear Feature Transformations for Noise Robust Speech Recognition*, Ph.D. Thesis, Institut de traitement des signaux (EPFL), Lausanne, Switzerland, 2004
- [27] Misra, H., Boulard, H., Tyagi, V., "New Entropy based Combination Rules In HMM/ANN Multi-stream ASR", In *Proc. ICASSP*, pp. 741-744, 2003.
- [28] Ellis, D.P., Singh, R., Sivasdas, S., "Tandem Acoustic Modeling in Large-vocabulary Recognition", In *Proc. ICASSP*, Vol. 1, pp. 517-520, 2001.
- [29] Sivasdas, S., Hermansky, H., "Hierarchical Tandem Feature Extraction", In *Proc. ICASSP*, 2002.
- [30] Sainath, T.N., Ramabhadran, B., Nahamoo, D., Kanevsky, D., Sethy, A., "Sparse Representation Features for Speech Recognition", In *Proc. Interspeech*, pp. 2254-2257, 2010.
- [31] Sivaram, G.S.V.S., Nemala, S.K., Elhilali, M., Tran, T.D., Hermansky, H., "Sparse Coding for Speech Recognition", In *Proc. ICASSP*, pp. 4346-4349, 2010.
- [32] Sivaram, G.S.V.S., Hermansky, H., "Sparse Multilayer Perceptron for Phoneme Recognition", *IEEE Trans. Audio, Speech, and Language Processing*, Vol. 20, No. 1, pp. 23-29, 2012.
- [33] Gemmeke, J.F., Virtanen, T., Hurmalainen, A., "Exemplar-based Sparse Representations for Noise Robust Automatic Speech Recognition", *IEEE Trans. Audio, Speech, and Language Processing*, Vol. 19, No. 7, pp. 2067-2080, 2011.
- [34] White, H., "Learning in Artificial Neural Networks: A Statistical Perspective", *Neural Computation*, Vol. 1, No. 4, pp. 425-464, 1989.
- [35] Zavaliagkos, G., Zhao, Y., Schwartz, R., Makhoul, J., "A Hybrid Segmental Neural Korea, pp. 612-615, 2004.
- [10] Okawa, S., Nakajima, T., Shirai, K., "A Recombination Strategy for multi-band Speech Recognition based on Mutual Information Criterion", In *Proc. Eurospeech*, Budapest, Hungary, pp. 603-606, 1999.
- [11] Motlicek, P., Cernocky, J., "Time-domain based Temporal Processing with Application of Orthogonal Transformations", In *Proc. Eurospeech*, Switzerland, pp. 821-824, 2003.
- [12] Athineos, M., Hermansky, H., Ellis, D., "LP-TRAP: Linear Predictive Temporal Patterns", In *Proc. ICSLP*, Korea, pp. 1154-1157, 2004.
- [13] Hermansky, H., "TRAP-TANDEM: Data-driven Extraction of Temporal Features from Speech", In *Proc. IEEE ASRU*, pp. 255-260, 2003.
- [14] Valente, F., Vepa, J., Plahl, C., Gollan, C., Hermansky, H., Schluter, R., "Hierarchical Neural Networks Feature Extraction for LVCSR System", In *Proc. InterSpeech*, Belgium, pp. 42-45, 2007.
- [15] Chen, B.Y., *Learning Discriminant Narrow band Temporal Patterns for Automatic Recognition of Conversational Telephone Speech*, Ph.D. Thesis, University of California, Berkeley, USA, 2005.
- [16] Hermansky, H., Ellis, D.P., Sharma, S., "Tandem Connectionist Feature Extraction for Conventional HMM Systems", In *Proc. ICASSP*, pp. 1635-1638, 2000.
- [17] Zhu, Q., Stolcke, A., Chen, B.Y., Morgan, N., "Using MLP Features in SRI's Conversational Speech Recognition System", In *Proc. InterSpeech*, pp. 2141-2144, 2005.
- [18] Valente, F., "Multi-stream Speech Recognition based on Dempster-Shafer Combination Rule", *Speech Communication*, Vol. 52, No. 3, pp. 213-222, 2010.
- [19] Kazemi, A.R., Sobhanmanesh, F., "MLP Refined Posterior Features for Noise Robust Phoneme Recognition", *Scientia Iranica, Trans. D: Computer Science & Engineering and Electrical Engineering*, Vol. 18, No. 6, pp. 1443-1449, 2011.
- [20] Park, J., Diehl, F., Gales, M.J.F., Tomalin, M., Woodland, P.C., "The Efficient Incorporation of MLP Features into Automatic Speech Recognition Systems", *Computer Speech and Language*, Vol. 25, No. 3, pp. 519-534, 2011.
- [21] Pinto, J., Garimella, S., Magimai-Doss, M., Hermansky, H., Boulard, H., "Analysis of MLP-Based Hierarchical Phoneme Posterior Probability Estimator", *IEEE Trans. Audio Speech Language Processing*, Vol. 19, No. 1, pp. 225-241, 2011.
- [22] Ikbali, S., Misra, H., Hermansky, H.,

- [42] Nejadgholi, I., Seyyedsalehi, S.A., "Nonlinear Normalization of Input Patterns to Speaker Variability in Speech Recognition Neural Networks", *Neural Computing and Applications*, Vol. 18, No. 1, pp. 45–55, 2009.
- [43] Shekofteh, Y., Almasganj, F., "Autoregressive Modeling of Speech Trajectory Transformed to the Reconstructed Phase Space for ASR Purposes", *Digital Signal Processing*, Vol. 23, No. 6, pp. 1923–1932, 2013.
- [44] Vali, M., Seyyedsalehi, S.A., "Robust Recognition of Telephone Speech using Proper Feature Extraction of Reverse Neural Networks", *IJECE*, Vol. 4, No. 1, pp. 21–29, 2008.
- [45] Shekofteh, Y., Almasganj, F., "Feature Extraction based on Speech Attractors in the Reconstructed Phase Space for Automatic Speech Recognition Systems", *ETRI Journal*, Vol. 35, No. 1, pp. 100–108, 2013.
- net/hidden Markov Model System for Continuous Speech Recognition", *IEEE Trans. Speech Audio Processing*, Vol. 2, No. 1, pp. 151–160, 1994.
- [36] Shekofteh, Y., Almasganj, F., "Improvement of Speech Recognition using Neural Net and Temporal Patterns", In *Proc. IKT2007*, pp. 1–8, 2007.
- [37] Chen C., Bilmes J., "MVA Processing of Speech Features", *IEEE Trans. Speech and Audio Processing*, Vol. 15, No. 1, pp. 257–270, 2007.
- [38] HTK (v.3.4), Hidden Markov Model Toolkit: <<http://htk.eng.cam.ac.uk/>>
- [39] Bijankhan, M., Sheikhzadegan, J., Roohani, M.R., Samareh, Y., Lucas, C., Tebyani, M. "FARSDAT-The Speech Database of Farsi Spoken Language", In *Proc. ACSST*, Vol. 2, pp. 826–830, 1994.
- [40] NOISEX-92, SPIB noise data, Available from: http://spib.rice.edu/spib/select_noise.html.
- [41] Kittler, J., Hatef, M., Duin, R.P.W., Matas, J., "On Combining Classifiers", *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 20, No. 4, pp. 226–239, 1998.

-
- 1 Automatic Speech Recognition
 - 2 Temporal Patterns
 - 3 Posterior Probability
 - 4 Linear Discriminant Analysis
 - 5 Principal Component Analysis
 - 6 Hierarchical
 - 7 Short Time Frames
 - 8 Data-Driven Temporal Processing
 - 9 Multi Layer Perceptron
 - 10 States of Phoneme
 - 11 Mel-Frequency Cepstral Coefficients
 - 12 Logarithm Filter Bank Energy
 - 13 Non-linguistic
 - 14 Co-articulation
 - 15 Sparse
 - 16 Mean Square Error
 - 17 Segment
 - 18 Tonotopic Multi Layer Perceptron
 - 19 Tonotopic
 - 20 Discrete Cosine Transform
 - 21 Real Time Factor

