بازشناسی مقاوم و توأم گفتار مستقیم و تلفنی با استخراج مناسب بردارهای بازنمایی و اصلاح آنها توسط معکوس‌سازی شبکه‌های عصبی

منصور ولی و سید علی سید صالحی

چکیده: در حال حاضر تلاش فراگیری برای طراحی سیستم‌های بازشناسی گفتار مقاوم نسبت به تنوعات گفتار صورت می‌گیرد. یکی از این تنوعات، گفتار تلفنی نسبت به گفتار مستقیم (تهیه شده در شرایط عاری از هر گونه نویز محیط) می‌باشد. در مقاله حاضر با بهره‌گیری از پارامترهای طیفی LHCB[1] و طراحی یک سری آزمایشهای عملی مشخص می‌گردد که این نوع بازنمایی برای طراحی سیستم‌های بازشناسی گفتار تلفنی و سیستم‌های بازشناسی توأم گفتار مستقیم و تلفنی که مبتنی بر شبکه‌های عصبی باشد نسبت به روش متداول MFCC مناسب‌تر است. سپس با استخراج بردارهای بازنمایی LHCB از گفتار مستقیم و تلفنی و طراحی مدل بازشناسی گفتار مبتنی بر شبکه عصبی MLP، یک سیستم بازشناسی توأم گفتار مستقیم و تلفنی ساخته می‌شود. آنگاه با استفاده از معکوس‌سازی شبکه‌های عصبی به روش گرادیان بردارهای بازنمایی گفتار تلفنی به سمت بردارهای بازنمایی گفتار مستقیم اصلاح می‌گردد و با تعلیم شبکه دیگری روی دادگان اصلاح شده تلفنی و دادگان مستقیم دست نخورده، افزایش ۱٫٤٪ در صحت بازشناسی گفتار تلفنی حاصل شده است. در مرحله بعد با استفاده از معکوس سازی عمومی شبکه‌های عصبی هر دو دسته بردارهای بازنمایی گفتار مستقیم و تلفنی به گونه‌ای اصلاح می‌شوند که بیشتر حاوی اطلاعات آوایی گفتار باشند و سایر تنوعات تا جای ممکن حذف شوند. با تعلیم شبکه دیگری روی این دادگان اصلاح شده افزایش ۲٫۹۸٪ در صحت بازشناسی گفتار تلفنی و ۱٫٦۸٪ در صحت بازشناسی گفتار مستقیم بدست آمده است.

کلید واژه: **بازشناسی‌مقاوم‌گفتار، بازنمایی، شبکه عصبی، معکوس‌سازی.**

۱- مقدمه

در سالهای اخیر کاربرد سیستم‌های بازشناسی گفتار روز به روز رو به افزایش است. بسیاری از سیستم‌های بازشناسی گفتار در کاربردهای محدود دقت بازشناسی بالایی دارند اما همین سیستم‌ها در کاربردهای واقعی و فراگیر دچار مشکل خواهند شد. یکی از این مشکلات وجود تنوعات[2] گفتار در شرایط متنوع صوتی است که این سیستم‌ها در آنها بکار می‌روند. این تنوعات ناشی از منابعی مثل نویز اضافه شونده، انواع کانال صوتی، مشخصات میکروفن، لهجه‌ها و شیوه‌های مختلف گویش و غیره می‌باشند. این عوامل باعث عدم تطبیق بین داده‌های تعلیم و دادگان آزمون شده، منجر به کاهش کیفیت بازشناسی خواهند شد [۱].

روشهای متعددی برای غلبه بر این عدم تطابق پیشنهاد شده است. این روشها شامل تنظیم پارامترهای مدل با یادگیری بدون سرپرستی یا با سرپرستی و یا روشهای پردازش سیگنال به منظور از بین بردن اثر نامطلوب نویز روی سیگنال می‌باشند [۲] و [۳]. گفتار تلفنی در مقایسه با گفتار مستقیم (تهیه شده در شرایط عاری از هرگونه نویز محیط) بسیاری از این تنوعات گفتار را در بردارد. در طراحی بسیاری از سیستم‌های بازشناسی گفتار تلفنی، برای بدست آوردن صحت بازشناسی مطلوب، مدل بازشناسی را با دادگان تلفنی تعلیم مجدد می‌دهند و به این ترتیب یک مدل جدید بازشناسی برای دادگان تلفنی نسبت به دادگان گفتار مستقیم بدست می‌آورند. اما تعلیم مجدد مدل برای دادگان تلفنی چه از نظر تهیه دادگان تلفنی و چه از نظر زمانی بسیار هزینه‌بر است. علاوه بر این از آنجایی که مدل روی دادگان تلفنی با پهنای باند محدود تعلیم می‌بیند بنابراین کیفیت بازشناسی نسبت به گفتار مستقیم پایین‌تر خواهد آمد [۴].

یک راه حل دیگر برای طراحی سیستم بازشناسی گفتار تلفنی، نگاشت غیرخطی از بردارهای بازنمایی گفتار تلفنی به بردارهای بازنمایی گفتار مستقیم است که آن را با استفاده از شبکه‌های عصبی MLP می‌توان محقق نمود اما در این روش اگرچه نیازی به طراحی مدل جدید بازشناسی گفتار تلفنی نمی‌باشد اما دقت بازشناسی نسبت به روشهای قبل بهبود چندانی ندارد. علاوه بر این به دلیل ضرورت همزمانی در سیگنال‌های گفتار تلفنی و مستقیم، این دو دسته دادگان باید همزمان ضبط شوند بنابراین نمی‌توان از مجموعه دادگان گفتاری مجزای تلفنی و مستقیم خود استفاده نمود [۵].

مطالعات روانشناسی نشان می‌دهد که انسان از دانش[3] قبلی خود برای یادگیری بهتر فعلی خود بهره می‌گیرد. علاوه بر این چندین وظیفه مرتبط را می‌تواند بطور توأمان یاد بگیرد [۶]. نقطه ضعف سیستم‌های بازشناسی گفتار در این است که آنها براساس دادگان تعلیمی خود یاد می‌گیرند و از قابلیتهای انسان در یادگیری بهره‌ای نمی‌برند. با دخیل کردن دانش قبلی در روند یادگیری، چندین خاصیت مطلوب مانند کاهش زمان یادگیری و تعمیم بهتر حاصل خواهد شد. انسان در مواجه شدن با یک مورد یادگیری جدید از دانشهای فراگرفته شده قبلی خود بهره می‌گیرد [۷]. به عنوان مثال به نظر می‌رسد کیفیت بالای بازشناسی گفتار تلفنی در انسان تحت

این مقاله در تاریخ ۹ خرداد ماه ۱۳۸۴ دریافت و در تاریخ ۳ آذر ماه ۱۳۸۴ بازنگری شد. این تحقیق توسط مرکز تحقیقات مخابرات براساس قرارداد شماره ۵۰۰/۱۴۵۲/ت پشتیبانی شده است.

منصور ولی، دانشکده مهندسی پزشکی، دانشگاه صنعتی امیرکبیر، خیابان حافظ، تهران، ایران (email: m_vali@bme.aut.ac.ir).

سید علی سیدصالحی ، دانشکده مهندسی پزشکی، دانشگاه صنعتی امیرکبیر، خیابان حافظ، تهران، ایران (email: ssalehi@aut.ac.ir).

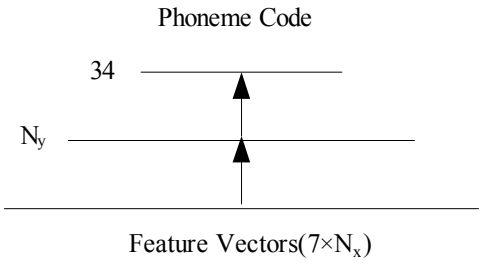
1. Logarithm of Energies in Hanning Critical Band Filter Banks
2. Variations
3. Knowledge

تأثیر دانش قبلی او از گفتار مستقیم است که قبلاً در محیط عاری از نویز فراگرفته شده است.

برای استفاده از دانش قبلی در یادگیری بهتر فعلی، توجه ویژه‌ای به تکنیکهای شبکه‌های عصبی صورت گرفته است [۸] و [۹]. با استفاده از شبکه‌های عصبی می‌توان با چند روش مختلف از دانش فراگرفته قبلی استفاده نمود؛ ازقبیل برنامه‌ریزی وزنهای شبکه [۸]، ایجاد مثالهای مجازی [۱۰] و استفاده از هدفهای خروجی اضافه که این تکنیک با عنوان مدل بازشناسی چند منظوره[1] شناخته می‌شود [۹] و [۱۱].

مقاله حاضر گامی در جهت بکارگیری دانش حاصل از تعلیم مدل روی گفتار مستقیم در بازشناسی بهتر گفتار مستقیم و تلفنی است. ما معتقدیم طراحی یک مدل واحد برای بازشناسی گفتار مستقیم و تلفنی باعث خواهد شد که در بازشناسی گفتار تخریب شدهٔ تلفنی بتوان از دانش نهفته در شبکه، ناشی از یادگرفتن گفتار مستقیم بهره گرفته و بردارهای بازنمایی گفتار تلفنی را بهبود بخشید. برای این منظور ابتدا لازم است بردارهای بازنمایی به گونه‌ای از گفتار استخراج شوند که برای دو نوع گفتار تلفنی و مستقیم سازگار باشند و امکان تعلیم توأمان آنها به یک مدل شبکه‌عصبی فراهم شود. در این تحقیق مشخص خواهد شد که بردارهای بازنمایی LHCB برای این منظور بهتر از MFCC عمل می‌نمایند. برای استفاده از دانش گفتار مستقیم در بهبود بازشناسی گفتار تلفنی از تکنیک بازشناسی چند منظوره در شبکه‌های عصبی استفاده شده است. به این ترتیب که بازشناسی آواهای گفتار، توأمان با تشخیص نوع کانال گفتار (تلفنی یا مستقیم) در یک شبکه MLP صورت می‌گیرد. سپس از دو روش معکوس‌سازی شبکه (معکوس سازی به روش گرادیان و معکوس سازی عمومی) استفاده شده و بردارهای بازنمایی در جهت حذف تنوعات از آنها اصلاح می‌شوند. آنگاه با تعلیم یک مدل شبکه عصبی دیگر روی این دادگان جدید، بازشناسی آواهای گفتار مستقیم و تلفنی با صحت بالاتری صورت می‌پذیرد.

در ادامه مقاله ابتدا در بخشهای ۲ و۳ به ترتیب ساختار شبکه‌های عصبی MLP و دادگان گفتاری استفاده شده در این تحقیق معرفی می‌شوند. سپس در بخش ۴ یک بحث تحلیلی روی دو نوع پارامتر بازنمایی گفتار LHCB و MFCC خواهد شد و نحوه استخراج بهینه بردارهای بازنمایی برای دادگان تلفنی بر اساس طراحی چند آزمون روی مدلهای شبکه عصبی به دست می‌آید. در بخش ۵ بر اساس یک سری آزمایش ثابت می‌گردد که بردارهای بازنمایی LHCB برای بازشناسی گفتار تلفنی و نیز بازشناسی‌های توأمان تلفنی و مستقیم نسبت به روش MFCC در مدلهای بازشناسی مبتنی بر شبکه‌عصبی ترجیح دارند. در بخش ۶ مقاله مبانی نظری معکوس‌سازی به روش گرادیان و معکوس‌سازی عمومی شبکه‌های عصبی بیان می‌گردد سپس در بخش ۷ با تعلیم توأمان شبکه روی بردارهای بازنمایی LHCB گفتار مستقیم و تلفنی، با استفاده از دو شیوه معکوس‌سازی به روش گرادیان و معکوس سازی عمومی، بردارهای بازنمایی در جهت افزایش صحت بازشناسی گفتار اصلاح می‌شوند. در پایان پس از بحث و تحلیل در بخش ۸، نتایج این تحقیق در بخش۹ جمع‌بندی می‌شوند. کلیه برنامه‌ها در محیط نرم‌افزار MATLAB پیاده‌سازی شده‌اند.



شکل ۱: مدل بازشناسی.

۲- ساختار مدل بازشناسی

مدل‌های بازشناسی بکار رفته در این تحقیق، مبتنی بر شبکه‌های عصبی MLP بوده و ساختار کلی آنها مشابه شکل ۱ می‌باشد. ساختار نورونی این شبکه بصورت $۷\times N_x$ - N_y - ۳۴ می‌باشد. که در آن N_x طول بردار بازنمایی است. بنابراین در ورودی شبکه بردار بازنمایی فریم جاری به همراه بردارهای بازنمایی ۳ فریم مجاور چپ و راست فریم جاری (در مجموع ۷ فریم) قرار می‌گیرند. در خروجی شبکه نیز به تعداد آواهای موجود در دادگان تعلیم یعنی ۳۴ نورون در نظر گرفته شده است.

N_y تعداد نورون‌های لایه مخفی شبکه است. انتخاب مقدار N_y بر اساس این قاعده است که تعداد الگوهای تعلیمی به شبکه باید ۴ تا ۱۰ برابر تعداد وزن‌های مجهول شبکه انتخاب شوند [۱۲]. کمتر از ۴ برابر خاصیت تعمیم شبکه را ضعیف کرده و اعتبار مدل را پایین می‌آورد و بزرگتر از ۱۰ برابر نیز منجر به بزرگی شبکه و طولانی شدن زمان تعلیم آن می‌گردد درحالیکه تأثیر ناچیزی در بهبود دقت بازشناسی می‌گذارد.

۳- دادگان گفتار

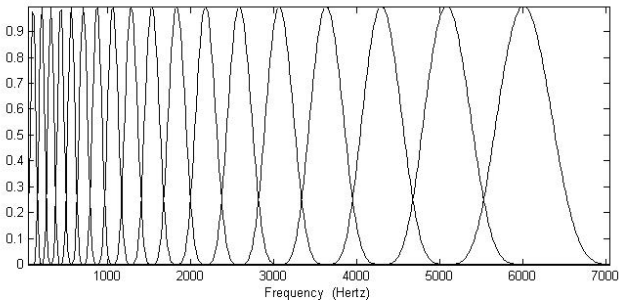
دادگان گفتاری مورد نیاز برای تعلیم شبکه‌ها از مجموعه دادگان فارس‌دات و فارس‌دات تلفنی انتخاب شده‌اند [۱۳]. این مجموعه دادگان شامل دو جمله مشترک بیان شده توسط ۶۴ گویشور تلفنی و ۲۰۰ گویشور مستقیم (غیرتلفنی) می‌باشد. این دو جمله تمام آواهای فارسی را در بر دارند اما طبیعی است که تمام ترکیب‌های آوایی را نخواهند داشت. بنابراین یک سیستم بازشناسی گفتار پیوسته مستقل از گوینده ولی وابسته به متن خواهد بود. علت این انتخاب به هدف ما برمی‌گردد که درحقیقت پیدا کردن روش‌های بهبود بردارهای بازنمایی با استفاده از قابلیتهای شبکه‌های عصبی است که وابسته به متن بودن دادگان به قدرت تعمیم نتایج این تحقیق لطمه‌ای وارد نمی‌کند. به این ترتیب از این مجموعه دادگان ۷۵٪ برای تعلیم شبکه‌ها و ۲۵٪ دیگر آن برای آزمون شبکه‌ها تخصیص داده شده‌اند. دو جمله مشترک انتخاب شده به صورت زیر می‌باشد.

"با روشن شدن هوا تظاهر کنندگان به سوی مجلس شورای ملی شروع به راهپیمایی کردند"

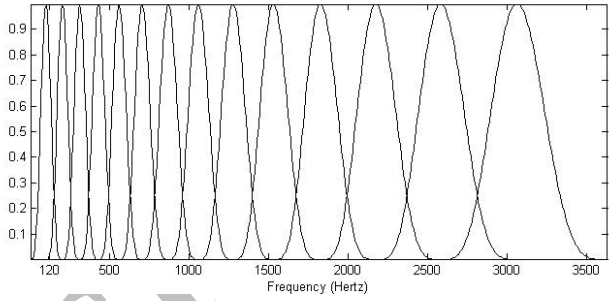
"مگر مژده اول چراغ قوه را خاموش نکرد"

نرخ نمونه برداری سیگنال گفتار برای گویشهای تلفنی ۸ کیلوهرتز و برای گویشهای غیرتلفنی برابر ۱۶ کیلوهرتز می‌باشد.

1. Multitask



شکل ۲: بانک فیلتر برای استخراج پارامترهای بازنمایی از گفتار مستقیم.



شکل ۳: بانک فیلتر برای استخراج پارامترهای بازنمایی از گفتار تلفنی.

٤- استخراج بازنمایی‌های LHCB و MFCC

٤-١- معرفی پارامترهای LHCB و MFCC

یکی از بهترین و رایجترین روشهای استخراج بازنمایی از گفتار مستقیم و تلفنی استخراج پارامترهای MFCC به همراه مشتقات اول و دوم آنها و اعمال هنجارسازی[1] طولی (کم کردن میانگین بردارهای بازنمایی هر گویش از هر بردار بازنمایی و تقسیم بر انحراف معیار آن بردارها) روی آنها می‌باشد [۱۴].

شکل ۲ یک بانک فیلتر ۱۸تایی نوع هنینگ که در محدوده فرکانسی ۸ کیلوهرتز توزیع شده است را نشان می‌دهد. محدوده صفر تا ۸ کیلوهرتز محدوده فرکانسی گفتار مستقیم است. پارامترهای LHCB لگاریتم انرژی گفتار در هر یک از این فیلترها می‌باشند که مجموعاً ۱۸ پارامتر انرژی می‌شود [۱۵] و پارامترهای MFCC با اعمال تبدیل کسینوسی روی پارامترهای LHCB در تعداد دلخواه بدست می‌آیند (معمولاً ۱۲ پارامتر در بازشناسی گفتار استخراج می‌گردد که با اضافه کردن لگاریتم انرژی کل فریم جمعاً ۱۳ پارامتر می‌گردد). بنابراین پارامترهای LHCB در حوزه طیف سیگنال گفتار هستند در حالی که پارامترهای MFCC در حوزه کپستروم می‌باشند.

برای گفتار با پهنای باند ۴ کیلو هرتز (گفتار تلفنی) تنها ۱۴ فیلتر از این ۱۸ فیلتر در محدوده فرکانسی آن واقع می‌شوند. این محدوده از بانک فیلتر در شکل ۳ نشان داده شده است. پارامترهای LHCB لگاریتم انرژی این ۱۴ فیلتر هستند. و پارامترهای MFCC با محاسبه تبدیل کسینوسی روی این ۱۴ پارامتر LHCB بدست می‌آیند. به این ترتیب ۱۳ پارامتر MFCC بدست آمده از گفتار با پهنای باند ۴ کیلوهرتز به دلیل همین تبدیل کسینوسی هیچ شباهتی به ۱۳ پارامتر بدست آمده از گفتار با پهنای باند ۸ کیلوهرتز ندارند. درحالی که انرژی ۱۴ فیلتر اول بانک (پارامترهای LHCB) برای هر دو گفتار مشابه همدیگر هستند.

1. Normalization

٤-٢- طراحی آزمایش برای تعیین محدوده فرکانسی مناسب در استخراج بردارهای بازنمایی از گفتار تلفنی

محدوده فرکانسی مناسب برای استخراج بازنمایی از گفتار تلفنی (دارای محدوده فرکانسی بین صفر تا ۴ کیلوهرتز) را معمولا بین ۱۲۵ تا ۳۷۰۰ هرتز انتخاب می‌کنند [۱۶]. زیرا که اثرات کانال‌های تلفنی مختلف در محدوده فرکانسهای پایین و بالای طیف ایجاد تخریب می‌کند. و این بخش از طیف گفتار تلفنی، اطلاعات مفیدی برای بازشناسی ندارد. اما از آنجایی که مدل بازشناسی در این تحقیق از نوع شبکه‌های عصبی MLP می‌باشد و شبکه این توانایی را دارد که مؤلفه‌های غیرمفید ورودی را با تخصیص وزنهای ضعیف به آنها کنار گذاشته و از مؤلفه‌های مفید ورودی استفاده کند؛ این سؤال مطرح می‌شود که آیا برای بازشناسی گفتار تلفنی در مدلهای بازشناسی مبتنی بر شبکه‌های عصبی نیز لازم است که قسمتهای ابتدایی طیف گفتار کنار گذاشته شود؟ به همین منظور آزمایش‌هایی به صورت زیر طراحی می‌شوند.

مدل بازشناسی بکار رفته در این آزمایش‌ها، مبتنی بر شبکه‌های عصبی MLP بوده و ساختار نورونی آنها مطابق شکل ۱ به صورت $N\times$ ۷۰-۷۰-۳۴ می‌باشد.

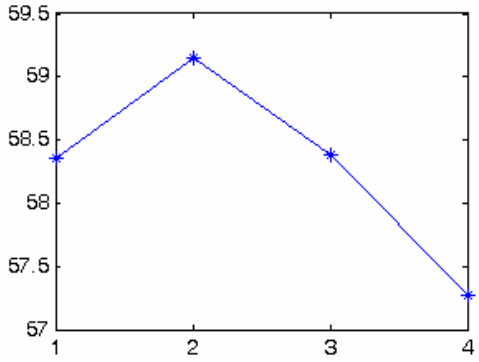
دادگان مورد نیاز برای تعلیم شبکه‌ها نیز مطابق بخش ۳ دو جمله مشترک بیان شده توسط تمام گویندگان تلفنی (۶۴ گویشور) استفاده شده است. از این مجموعه دادگان ۴۸ گوینده برای تعلیم و ۱۶ گوینده برای آزمون استفاده شده‌است.

٤-٢-١- بررسی تأثیر انرژی فیلترهای فرکانس پایین بانک در دقت بازشناسی گفتار تلفنی

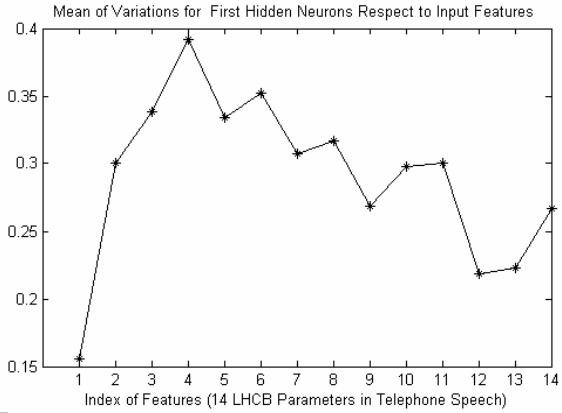
مطابق شکلهای ۲ و ۳ از ۱۸ بانک فیلتر طراحی شده برای استخراج پارامترهای LHCB یا MFCC از گفتار مستقیم، ۱۴تای ابتدایی آن در محدوده طیف گفتار تلفنی قرار می‌گیرند. هدف این است که از میان بردار ۱۴ مؤلفه‌ای انرژی فیلترها، تأثیر چند مؤلفه اول آن در صحت بازشناسی آواهای گفتار تلفنی به دست آید. برای این منظور در مرحله اول یک مدل بازشناسی گفتار مبتنی بر شبکه عصبی توسط بردارهای بازنمایی ۱۴ مؤلفه‌ای LHCB تعلیم داده می‌شود (انتخاب $N=۱۴$). مرحله دوم با حذف مؤلفه اولِ این بردارهای بازنمایی، شبکه عصبی دوم تعلیم داده می‌شود (انتخاب $N=۱۳$) و در مرحله سوم با حذف دو مؤلفه اول بردارهای بازنمایی، شبکه عصبی سوم تعلیم داده می‌شود (انتخاب $N=۱۲$) و بالاخره در مرحله چهارم با حذف سه مؤلفه اول بردارهای بازنمایی، شبکه چهارم تعلیم داده می‌شود (انتخاب $N=۱۱$). نتایج آزمون این چهار شبکه در جدول ۱ آمده است. از روی نتایج جدول ۱ که جهت درک بهتر در شکل ۴ نیز ترسیم شده‌اند؛ مشخص است که حذف انرژی فیلتر اول از بردارهای بازنمایی تلفنی تأثیر مثبت در صحت بازشناسی آوا داشته است ولی انرژی فیلترهای دوم و سوم بردارهای بازنمایی نباید کنار گذاشته شوند و حاوی اطلاعات مفیدی برای بازشناسی آواها می‌باشند. طبیعتاً نتیجه این آزمون برای استخراج پارامترهای MFCC از گفتار تلفنی نیز قابل تعمیم است. چرا که پارامترهای MFCC هم تبدیل کسینوسی همین بردارهای LHCB هستند.

٤-٢-٢- تحلیل حساسیت شبکهٔ بازشناسی گفتار تلفنی نسبت به انرژی فیلترهای بانک

برای حذف انرژی فیلتر اول از بردارهای بازنمایی علاوه بر روش فوق، بر اساس روش تحلیلی زیر نیز می‌توان آن را اثبات کرد. در این روش یک



شکل ۴: صحت بازشناسی گفتار تلفنی مربوط به چهار شبکه فوق.



شکل ۵: میزان حساسیت نورونهای لایه مخفی شبکه به ۱۴ مؤلفه بردار LHCB در ورودی شبکه.

جدول ۱: نتایج حذف پارامترهای فرکانس پایین از پارامترهای LHCB.

بردار بازنمایی	صحت بازشناسی آواها
۱۴ پارامتر LHCB	۵۸٫۳۵
حذف پارامتر اول	۵۹٫۱۵
حذف دو پارامتر اول	۵۸٫۳۸
حذف سه پارامتر اول	۵۷٫۲۷

دید فراگیر نسبت به میزان تأثیر همه مؤلفه‌های بردار ورودی در عملکرد شبکه می‌توان پیدا کرد.

اگر در شبکه اول تعلیم داده شده روی بردارهای ۱۴ تایی LHCB، برای ۷۰ نورون لایه مخفی شبکه، میزان حساسیت خروجی این نورونها را نسبت به ورودیهای شبکه محاسبه کنیم یعنی $|\partial y_j / \partial x_i|$ و این مقدار را روی تمامی دادگان تعلیم شبکه میانگین گیری کنیم خواهیم داشت

$$S_{ji} = \sum_{\forall x_i} \left| \frac{\partial y_j}{\partial x_i} \right| \qquad j = ۱,۲,\ldots,۷۰ \qquad i = ۱,۲,\ldots,۱۴ \qquad (۱)$$

اکنون اگر بخواهیم میزان حساسیت نورونهای لایه مخفی شبکه را نسبت به هر یک از این ۱۴ مؤلفه ورودی محاسبه کنیم کافی است S_{ji}ها را روی ۷۰ نورون لایه مخفی میانگین‌گیری کنیم یعنی

$$S_i = \frac{۱}{۷۰} \sum_{j=۱}^{۷۰} S_{ji} \qquad i = ۱,۲,\ldots,۱۴ \qquad (۲)$$

بردار ۱۴ تایی S نمایانگر حساسیت نورونهای لایه مخفی شبکه به ۱۴ مؤلفه بردار بازنمایی ورودی است که در شکل ۵ نشان داده شده است. بر طبق این نمودار واضح است که مؤلفه اول بردار بازنمایی تلفنی تاثیر بسیار کمی در تعیین وضعیت نورونهای شبکه دارد و می‌توان آنرا حذف کرد.

علت کوچک شدن مقدار حساسیت برای این مؤلفه از بردار بازنمایی را می‌توان اینگونه تحلیل کرد که شبکه در طول تعلیم به دلیل مؤثر نبودن این ورودی در ایجاد تمایز بین آواها، اثر آن را با وزنهای ضعیف‌تری که به آن تخصیص می‌دهد کاهش خواهد داد. بنابراین لازم است این مؤلفه از بردار بازنمایی حذف شود تا این بار اضافی از دوش شبکه برداشته شود. بنابراین در ادامه تحقیق برای استخراج بازنماییهای LHCB و MFCC از گفتار تلفنی، انرژی فیلتر اول بانک تاثیر داده نمی‌شود و فیلتر چهاردهم نیز تا محدوده ۳۵۵۰ هرتز را پوشش داده است.

۴-۳- ساختن بردارهای بازنمایی با اضافه کردن مشتقات اول و دوم پارامترها به آنها

از آنجایی که در انتخاب بهینه بردارهای بازنمایی گفتار از مشتقات اول و دوم پارامترها در کنار خود پارامترها نیز استفاده می‌شود بنابراین بردارهای بازنمایی MFCC به همراه مشتقات اول و دوم آنها برای هر دو نوع گفتار مستقیم و تلفنی شامل ۳۹ مؤلفه می‌گردد (۳۹=۱۳+۱۳+۱۳ مؤلفه). اضافه کردن مشتقات اول و دوم پارامترهای LHCB به آنها به صورت زیر انجام می‌شود. برای گفتار مستقیم یک بردار بازنمایی شامل ۱۸ پارامتر LHCB و مشتقات اول و دوم آنها که جمعاً شامل ۵۴ پارامتر می‌باشد ساخته می‌شود. اما از آنجایی که هدف نهایی تعلیم یک شبکه واحد بازشناسی گفتار مستقیم و تلفنی است باید طول بردارهای بازنمایی گفتار مستقیم و تلفنی با هم برابر باشند. به همین جهت در ساختن بردار بازنمایی گفتار تلفنی علاوه بر اینکه اولین مؤلفه از بردار ۱۴تایی LHCB تلفنی صفر قرار داده می‌شود (دلیل آن در بخش ۴-۲ آمده است) ۴ مؤلفه صفر که مربوط به فیلترهای محدوده ۳٫۶ تا ۸ کیلوهرتز است نیز به انتهای آن اضافه شده و یک بردار ۱۸ تایی از پارامترهای LHCB تلفنی ساخته می‌شود. سپس با اضافه کردن مشتقات اول و دوم این بردار به آن، بردار بازنمایی به طول ۵۴ مشابه بردار بازنمایی گفتار مستقیم بدست می‌آید. حال نه تنها طول بردار بازنمایی گفتار مستقیم و تلفنی برابر همدیگر هستند بلکه مؤلفه‌های نظیر هم از بردارهای بازنمایی گفتار مستقیم و تلفنی، مربوط به فیلترهای نظیر هم در بانک فیلتر هستند و همین باعث تشابه این دو بردار بازنمایی شده که امکان تعلیم توأمان به یک شبکه را فراهم می‌نماید.

۵- مقایسه بردارهای بازنمایی MFCC و LHCB در بازشناسی گفتار مبتنی بر شبکه‌های عصبی MLP

اکنون دو دسته بردار بازنمایی MFCC (۳۹ مؤلفه‌ای) و LHCB (۵۴ مؤلفه‌ای) روی مدلهای بازشناسی گفتار تلفنی، گفتار مستقیم و بازشناسی توأمان گفتار مستقیم و تلفنی بررسی شده و در عمل کاربرد این دو دسته بازنمایی ارزیابی می‌شوند.

توجه به این نکته ضروری است که اگرچه طول بردارهای بازنمایی MFCC کمتر از LHCB می‌باشد و ظاهراً مقایسه بین دو نوع بردار بازنمایی با طول غیریکسان منطقی به نظر نمی‌رسد ولی همانگونه که در بخش ۴-۱ مطرح شد پارامترهای ۱۳ تایی MFCC از تبدیل کسینوس روی پارامترهای ۱۸ تایی LHCB به‌دست می‌آیند. بنابراین تعداد فیلترها در بانک فیلتر اعمال شده روی طیف سیگنال، برای هر دو نوع بازنمایی یکسان است. به عبارتی دیگر می‌توان گفت محتوای اطلاعاتی این دو نوع بازنمایی یکسان، اما نحوه بیان آنها متفاوت هستند. به این ترتیب مقایسه بین کارآیی آنها منطقی است.

جدول ۲: صحت بازشناسی آواهای غیرسکوت، برای دادگان آزمون دو نوع بازنمایی

مدل بازشناسی		MFCC	LHCB
گفتار مستقیم		۸۲٫۴۸	۸۲٫۶۷
گفتار تلفنی		۶۴٫۱۸	۶۷٫۱۲
توأمان	مستقیم	۷۹٫۷۸	۸۱٫۹۲
گفتار	تلفنی	۶۱٫۱۱	۶۸٫۳۹

۵-۱- مقایسه بردارهای بازنمایی MFCC و LHCB در بازشناسی گفتار مستقیم و تلفنی

برای این منظور سه دسته آزمایش طراحی شده است. مدل بازشناسی بکار رفته در تمام آزمایشها، مبتنی بر شبکه‌های عصبی MLP مشابه شکل ۱ می‌باشد. ساختار نورونی این شبکه برای بردارهای بازنمایی LHCB بصورت ۳۴−۷۰−۷×۵۴ و برای بردارهای بازنمایی MFCC بصورت ۳۴−۷۰−۷×۳۹ می‌باشد. برای این قسمت از تحقیق از دو جمله مشترک بیان شده توسط تمام گویندگان تلفنی (۶۴ گویشور) و گویندگان غیرتلفنی (۱۰۰ گویشور) استفاده شده است. از هر مجموعه دادگان ۷۵٪ برای تعلیم شبکه‌ها و ۲۵٪ دیگر آن برای آزمون شبکه‌ها تخصیص داده شده است. دو جمله مشترک انتخاب شده همان جملاتی هستند که در بخش ۳ معرفی شدند.

آزمایش اول: تعلیم بردارهای بازنمایی MFCC و LHCB گفتار مستقیم به دو شبکه مجزا.

آزمایش دوم: تعلیم بردارهای بازنمایی MFCC و LHCB گفتار تلفنی به دو شبکه مجزا.

آزمایش سوم: تعلیم بردارهای بازنمایی MFCC مستقیم و تلفنی به یک شبکه و بردارهای بازنمایی LHCB مستقیم و تلفنی به یک شبکه دیگر.

بر اساس آزمایشهای مذکور نتایج صحت بازشناسی آواها برای دادگان آزمون این دو نوع بازنمایی در جدول ۲ آمده است. صحت بازشناسی بر اساس درصد صحت فریم‌های غیرسکوت بیان شده است. زیرا که تعداد فریم‌های سکوت گفتار از سایر آواها بیشتر بوده و صحت بازشناسی آنها برای هر نوع بازنمایی بسیار بالا است. و مانع از مشهود شدن تفاوت روش‌های مختلف بازشناسی گفتار و انواع بهبود در بازنمایی که مد نظر این مقاله است می‌گردد.

نتایج آزمایش اول نشان می‌دهد که بازشناسی گفتار مستقیم برای این دو دسته بازنمایی تفاوت چندانی با هم ندارند و از آنجایی که طول بردار بازنمایی MFCC از طول بردار بازنمایی LHCB کوتاه‌تر است برای بازشناسی گفتار مستقیم، پارامترهای MFCC ترجیح خواهند داشت.

در آزمایش دوم نتایج بهتر بازنمایی LHCB نسبت به MFCC برای بازشناسی گفتار تلفنی، گویای آن است که بازنمایی LHCB برای دادگانِ تخریب شده و نویزی مثل گفتار تلفنی مناسب‌تر است.

مقایسه نتایج حاصل از آزمایش سوم درمقایسه با نتایج آزمایش‌های اول و دوم نشان می‌دهد که بکارگیری بردارهای بازنمایی MFCC در شبکه بازشناسی توأمان گفتار مستقیم و تلفنی، نسبت به شبکه‌های بازشناسی مجزای گفتار مستقیم و تلفنی نتایج ضعیف‌تری می‌دهد. زیرا که بازنمایی‌های MFCC برای گفتار مستقیم و تلفنی تطابق منطقی با هم ندارند. بنابراین در طراحی سیستم‌های بازشناسی گفتار توأمان تلفنی و مستقیم نمی‌تواند کاربرد داشته باشد. اما استفاده از بازنماییهای طیفی گفتار مانند پارامترهای LHCB به دلیل تطابق منطقی بازنماییهای نظیر از آواهای گفتار مستقیم و تلفنی با همدیگر، امکان طراحی سیستم‌های بازشناسی توأمان گفتار مستقیم و تلفنی را فراهم می‌نماید. علاوه بر این از روی نتایج جدول ۲ می‌توان مشاهده کرد که شبکه بازشناسی توأم گفتار مستقیم و تلفنی برای بازنمایی‌های LHCB تلفنی، حدود ۱٫۲ درصد صحت بالاتری نسبت به شبکه بازشناسی خالص تلفنی (آزمایش دوم) دارد. زیرا دانش نهفته از تعلیم دادگان گفتار مستقیم در شبکه، باعث رشد بازشناسی گفتار تلفنی شده است. البته برای گفتار مستقیم نتیجه بازشناسی نسبت به آزمایش اول کمی پایین‌تر بدست آمده است اما در هر حال به رجحان داشتن بازنمایی‌های LHCB نسبت به MFCC لطمه‌ای نمی‌زند.

۵-۲- نتایج حاصل از آزمونها

بر اساس بحثهای استدلالی و آزمایش‌های مطرح شده در مجموع نتایج زیر بدست می‌آید.

الف. استخراج بازنمایی MFCC به همراه هنجارسازی مناسب آنها یکی از بهترین روشهای استخراج بازنمایی برای گفتار مستقیم است.

ب. برای مدلهای بازشناسی گفتار تلفنی مبتنی بر شبکه‌های عصبی به دلایلی که در متن بحث شد استخراج بازنماییهای طیفی مانند LHCB نتایج بهتری نسبت به بازنماییهای حوزه کپستروم (MFCC) دارد.

ج. در مدلهای بازشناسی تلفنی-مستقیم مبتنی بر شبکه‌های‌عصبی نیز به دلیل ضرورت سازگاری بین مؤلفه‌های بردار بازنمایی در گفتار تلفنی و مستقیم باز هم بازنماییهای طیفی مانند LHCB بر بازنماییهای حوزه کپستروم (MFCC) ترجیح پیدا می‌کنند.

نتیجه سوم برای ما حاوی نکته ارزشمندی است که می‌توان با طراحی مدلهای بازشناسی توأم گفتار مستقیم و تلفنی نه تنها کمبود دادگان جمع آوری شده تلفنی را از طریق دادگان مستقیم تأمین نمود بلکه با بکارگیری روشهای تطبیقی و نیز شیوه‌های معکوس سازی شبکه‌های عصبی بردارهای بازنمایی تلفنی را به سمت بردارهای بازنمایی مستقیم اصلاح کرد و نهایتاً به سیستم‌های بازشناسی گفتار مستقیم-تلفنی که نتایج بهتری در بازشناسی گفتار تلفنی و هم گفتار مستقیم در مقایسه با سیستم‌های مجزای بازشناسی گفتار تلفنی یا مستقیم دارند دسترسی پیدا کرد. در ادامه ایده فوق مورد بررسی و آزمون قرار می‌گیرد.

۶- معکوس‌سازی شبکه‌های عصبی

الگوریتم‌های معکوس‌سازی شبکه‌های عصبی را می‌توان به دو دسته اصلی تقسیم کرد:

۱) جستجوی کامل[1]

۲) جستجوی تک‌عنصری[2]

روشهای جستجوی کامل معمولاً زمانی که ابعاد و محدوده تغییرات هر یک از متغیرهای ورودی محدود باشد کاربرد دارد. این روش به دنبال پیدا کردن مجموعه جوابهای ممکن در ورودی شبکه است که منجر به یک خروجی مطلوب می‌گردند. در روشهای جستجوی تک عنصری در هر بار جستجو یک جواب ممکن یا یک جواب خاص از این مجموعه جوابها

1. Exhaustive Search
2. Single Element Search

بدست می‌آید که در برخی از روشهای آن وابسته به مقادیر اولیه شروع الگوریتم است [۱۷].

٦-۱- جستجوی تک‌عنصری به روش گرادیان

توجه به معکوس‌سازی تک‌عنصری شبکه‌های عصبی از طریق گرادیان[1] اولین بار توسط ویلیامز [۱۸] مطرح شد. این روش‌ها با استفاده از روش استاندارد بهینه‌سازی با پس‌انتشار خطا برای اصلاح یک ورودی اولیه کار می‌کنند. جستجو با یک بردار ورودی x^* شروع می‌شود. اگر x_k^t، k امین مؤلفه بردار x^t باشد، با هدف کاهش گرادیان، رابطه بازگشتی زیر به دست می‌آید

$$x_k^{t+1} = x_k^t - \eta \frac{\partial E}{\partial x_k^t} \qquad (۳)$$

که η طول گام و t اندیس تکرار است. با فرض یک توپولوژی جلوسوی عمومی، تکرار برای معکوس‌سازی در (۳) می‌تواند به صورتی که در ادامه می‌آید حل شود: با فرض $\delta_k = \partial E / \partial x_k$ و $k \in I$ که برای هر نورون

$$\delta_j = \begin{cases} \varphi_j'(o_j)(o_j - t_j) & j \in O \\ \varphi_j'(o_j) \sum_{m \in H,O} \delta_j w_{jm} & j \in I, H \end{cases} \qquad (۴)$$

I و H و O: به ترتیب مجموعه‌های نورون‌های ورودی، لایهٔ مخفی و خروجی ؛

w_{jm} : مقدار وزن از نورون jبه نورون m ؛

φ_j' : مشتق تابع غیرخطی نورون jام ؛

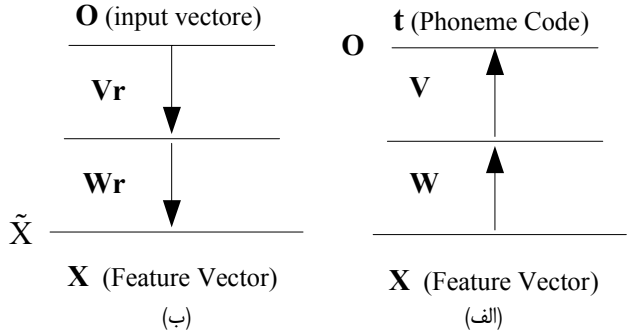
o_j : میزان فعالیت نورون jام ؛

t_j : خروجی مطلوب برای نورون jام .

مشابه روش استاندارد پس‌انتشار خطا، مشتقات نورون، δ_jها باید با ترتیب برعکس از خروجی به ورودی حل شوند. به این ترتیب مشتق خطای خروجی نسبت به بردارهای ورودی شبکه، مطابق رابطه بازگشتی (۴) محاسبه شده و این روند آنقدر تکرار می‌شود تا مقدار خطای خروجی یا شیب آن از یک حد آستانه کمتر شود. عدم وجود فیدبک، تنها فرض لازم برای اتصالات نورون‌ها در (۴) است. در ادامه از این روش معکوس‌سازی شبکه‌های عصبی، با نام روش معکوس‌سازی گرادیان یاد خواهد شد.

٦-۲- جستجوی تک‌عنصری به روش تعلیم شبکه معکوس

در این روش پس از تعلیم شبکه مستقیم، مطابق شکل ۶-الف به‌ازای بردار بازنمایی ورودی X مقدار هدف، بردار t (کد آوایی نظیر X) می‌باشد. اما مقدار خروجی واقعی شبکه پس از تعلیم، بردار O می‌شود و خطای شبکه برای این ورودی، $|t-O|$ است. حال شبکه MLP دیگری مطابق شکل ۶-ب به عنوان شبکه معکوس تعلیم داده می‌شود که ورودیهای این شبکه خروجیهای شبکه مستقیم (بردار O) و اهداف خروجی این شبکه بردارهای بازنمایی است که در ورودی شبکه شکل ۶-الف مورد استفاده قرار گرفته‌اند (بردار بازنمایی X). در حالی که



شکل ۶: (الف) شبکه MLP مستقیم و (ب) شبکه MLP معکوس.

خروجی واقعی این شبکه نیز بردار $\tilde{X}$ خواهد شد و خطای خروجی برای آن نیز برابر $|X - \tilde{X}|$ می‌گردد. نحوه محاسبه خروجی شبکه و خطای تعلیم در روابط (۵) و (۶) آمده است. روش تعلیم شبکه معکوس شکل ۶-ب مشابه روش تعلیم شبکه مستقیم (شکل ۶-الف) است. به این ترتیب که خطای خروجی شبکه نسبت به اهداف خروجی برخلاف جهت گرادیان پس‌انتشار شده، وزنهای شبکه اصلاح می‌گردند. این روند تا آنجا که صحت بازشناسی روی دادگان آزمون به ماکزیمم مقدار خود یا خطای خروجی پایین‌تر از یک حد آستانه برسد ادامه پیدا می‌کند.

$$O = f(V \times f(W \times X)) \qquad \tilde{X} = f(W_r \times f(V_r \times O)) \qquad (۵)$$

$$E = |t - O| \qquad E = |X - \tilde{X}| \qquad (۶)$$

اگرچه خطای نهایی تعلیم در این شبکه نسبت به شبکه مستقیم بزرگتر است اما به میانگین تقریبی مجموعه جوابهای ممکن منتهی می‌گردد [۱۱]. به همین جهت هم در ادامه از این روش با نام روش معکوس سازی عمومی یاد خواهد شد. در ادامه از این دو روش معکوس سازی شبکه‌های عصبی برای اصلاح بردارهای بازنمایی گفتار استفاده می‌شود.

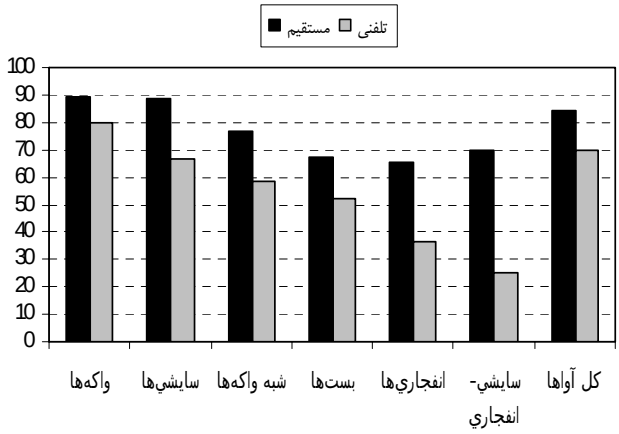
۷- اصلاح بردارهای بازنمایی با استفاده از معکوس سازی شبکه‌های عصبی

۷-۱- انتخاب بردار بازنمایی و دادگان

همانگونه که در بخش ۵ نشان داده شد بردارهای بازنمایی طیفی مانند LHCB برای گفتار مستقیم و تلفنی و نیز برای انواع گفتار تلفنی تطابق بهتری دارند. بنابراین در این بخش از تحقیق که هدف ما ارتقاء کیفیت بازشناسی گفتار در سیستم بازشناسی توأم گفتار مستقیم و تلفنی است بردارهای بازنمایی LHCB به همراه مشتقات اول و دوم آنها که در مجموع طولی برابر ۵۴ مؤلفه پیدا می‌کنند مورد استفاده قرار می‌گیرند و قبلاً در بخش ۴-۳ بیان شد که نه تنها طول این بردارهای بازنمایی، برای گفتار مستقیم و تلفنی برابر همدیگر هستند بلکه مؤلفه‌های نظیر هم از این دو نوع بردار بازنمایی، مربوط به فیلترهای نظیر هم در بانک فیلتر هستند و تشابه این دو بردار بازنمایی در مؤلفه‌های غیرصفر، امکان تعلیم توأمان به یک شبکه را فراهم می‌نماید.

دادگان استفاده شده برای این بخش از تحقیق، دادگانی است که در بخش ۳ مقاله به صورت کامل معرفی شده است.

1. Gradient Descent Method



شکل ۷: درصد بازشناسی دسته‌های آوایی مختلف در دادگان گفتار تلفنی و مستقیم.

۷-۲- مدل بازشناسی مرجع

در اولین مرحله یک مدل مرجع برای بازشناسی توأمان گفتار مستقیم و تلفنی مبتنی بر شبکه‌های عصبی MLP مطابق شکل ۱ طراحی می‌گردد. ساختار نورونی این شبکه بصورت ۳۴-۱۰۰-۵۴×۷ می‌باشد. خروجی شبکه دارای توابع غیرخطی از نوع تانژانت هایپربولیک[1] هستند. انتخاب این نوع تابع غیرخطی بجای تابع غیرخطی سیگموئید جهت تسریع در همگرایی شبکه است. مقادیر وزنهای اولیه شبکه بصورت تصادفی و در محدوده ۱- تا ۱ در نظر گرفته شده‌اند و ضریب یادگیری در شروع تعلیم برابر ۰/۰۰۱ و در حین تعلیم در هر تکرار با ضریب ۰/۹۵ کاهش می‌یابد که باعث تعلیم بهینه شبکه می‌گردد. ضریب ممنتم[2] نیز برابر ۰/۲ در نظر گرفته شده است.

شبکه مذکور چندین بار با مقادیر وزنهای تصادفی متفاوت اولیه تعلیم داده می‌شود و نهایتاً صحت بازشناسی گفتار تلفنی برابر ۰/۱±۶۹/۷۶ درصد و صحت بازشناسی گفتار مستقیم برابر ۰/۲±۸۴/۱ درصد بدست آمده است. صحت بازشناسی بر حسب درصد صحت فریم‌های غیرسکوت بیان شده است. به این ترتیب انحراف معیار در بازشناسی دادگان تلفنی برابر ۰/۱٪ و برای دادگان گفتار مستقیم برابر ۰/۲٪ می‌باشد. این مقادیر میزان اعتبار مدل بکار رفته در این تحقیق را نشان می‌دهد.

نمودار میله‌ای شکل ۷ درصد صحت بازشناسی آواها برای شش دسته مختلف آواهای گفتار یعنی واکه‌ها، سایشی‌ها، شبه واکه‌ها، بست‌ها، انفجاری‌ها و سایشی-انفجاری‌ها را به تفکیک نشان می‌دهد. نکات قابل توجه این نمودار بصورت زیر است:

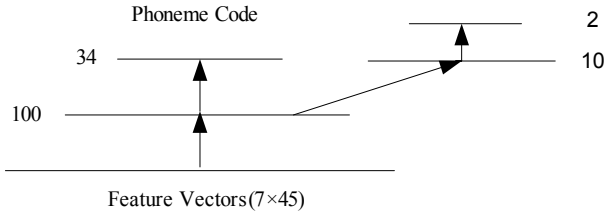
الف: بیشترین صحت بازشناسی برای هر دو دسته دادگان در واکه‌های گفتار حاصل شده است.

ب: بیشترین اثر تخریبی کانال تلفنی بر روی آواهای انفجاری و سایشی-انفجاری واقع شده است.

۷-۳- اصلاح بردارهای بازنمایی با استفاده از معکوس سازی گرادیان

در این روش ابتدا یک شبکه عصبی واحد برای بازشناسی گفتار مستقیم و تلفنی مشابه شبکه‌عصبی مدل مرجع تعلیم داده می‌شود. با این تفاوت که یک شاخه اضافه به لایه پنهان شبکه اضافه می‌گردد که خود

1. Hyperbolic
2. Momentum



شکل ۸: ساختار مدل بازشناسی دو منظوره.

شامل یک لایه پنهان ۱۰ نورونی و دو نورون در خروجی است که برای تشخیص نوع کانال (مستقیم یا تلفنی) دو بیت تخصیص داده شده است. این یک مدل بازشناسی دو منظوره است؛ بازشناسی آوا و بازشناسی نوع کانال که ساختار آن در شکل ۸ نشان داده شده است. نتایج آزمون این شبکه به همراه نتایج آزمون شبکه مرجع در جدول ۳ آورده شده است.

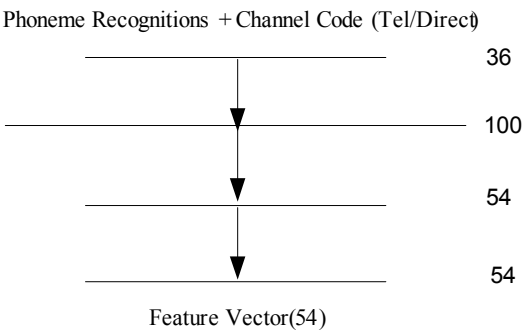
نتایج نشان می‌دهد که این شبکه با صحت بسیار بالایی نوع کانال تلفنی را از کانال گفتار مستقیم تمایز می‌دهد ۱۰۰٪-۹۹/۹۵٪ و افت ناچیز در بازشناسی آواهای گفتار مستقیم ۰/۱۸٪ و تلفنی ۰/۲۷٪ نسبت به شبکه مرجع کاملاً قابل توجیه است. زیرا شاخه اضافه شده برای شناسایی نوع کانال گفتار، در این شبکه نسبت به شبکه مرجع یک بار اضافه‌ای است که بر آن تحمیل شده است. به عبارتی وزنهای لایه اول این شبکه باید برای برآورده شدن دو هدف (شناسایی درست آواها و تشخیص درست نوع کانال) بهینه شوند.

حال بردارهای بازنمایی تلفنی به ورودی شبکه داده شده و کد دوبیتی خروجی شبکه (مربوط به تشخیص کانال گفتار) بصورت تحمیلی برابر کد کانال گفتار مستقیم قرار داده می‌شود و خطای خروجی مطابق روابط (۳) و (۴) به منظور اصلاح ورودی پس‌انتشار می‌شود. نکاتی که در اعمال این روش پس‌انتشار بر روی ورودی شبکه در نظر گرفته شده است به صورت زیر می‌باشد:

۱. خروجی مطلوب شبکه برای ۳۴ نورون شناسایی کننده آواها همان کد باینری آوای شناسایی شده توسط شبکه فرض می‌شود و خطای ناشی از این کد و خروجی واقعی شبکه نیز برای اصلاح ورودی پس‌انتشار می‌شود. به این ترتیب ورودی شبکه نه تنها در جهت تشخیص نوع کانال مستقیم اصلاح می‌گردد بلکه اطلاعات آوایی قبلی‌اش را نیز حفظ می‌کند.
۲. از میان هفت فریم‌متوالی که در ورودی شبکه قرار می‌گیرند تنها فریم میانی اصلاح می‌گردد. ولی تمام فریمهای دادگان تلفنیِ تعلیم و آزمون به این شیوه اصلاح می‌شوند.
۳. پس از اتمام الگوریتم اصلاح دادگان تلفنی، این دادگان اصلاح شده تلفنی به ورودی مدل بازشناسی دو منظوره داده شد و ۹۷/۸٪ از این دادگان از نوع گفتار مستقیم شناسایی شد و صحت بازشناسی آوا نیز تقریباً ثابت ماند. این نتیجه حاکی از موفقیت در اصلاح دادگان تلفنی در جهت نزدیک شدن به دادگان گفتار مستقیم است.

پس از اتمام مراحل فوق با رسم بردارهای بازنمایی تلفنیِ اصلاح شده بر روی بردارهای اولیهٔ نظیر آنها متوجه این نکته شدیم که بیشترین اصلاح و تغییر در مؤلفه‌های صفر بردارهای بازنمایی اتفاق افتاده است.

به عبارتی دیگر، مؤلفه‌هایی از بردارهای بازنمایی تلفنی که نسبت به مؤلفه‌های نظیر از بردارهای بازنمایی گفتار مستقیم وجود ندارند به نوعی از روی دانش نهفته در مدل بازشناسی دو منظوره تخمین زده شده‌اند.



شکل ۹: ساختار شبکه عصبی معکوس.

حال که بردارهای بازنمایی گفتار مستقیم و تلفنی به هم نزدیک شده‌اند. یک شبکه جدید بازشناسی آوا مشابه شبکه مرجع روی دو دسته از بردارهای بازنمایی گفتار مستقیم و تلفنی اصلاح شده تعلیم داده می‌شود. نتایج آزمون این شبکه در جدول (۳) با عنوان معکوس بروش گرادیان آمده است. به این ترتیب نزدیک به ۱/۴٪ افزایش در صحت بازشناسی گفتار تلفنی بدست آمده است. در بازشناسی گفتار مستقیم با توجه به انحراف معیار ۰/۲٪ که در نتیجه بازشناسی شبکه مرجع داشتیم، رشدی بدست نیامده که این نتیجه با توجه به اینکه هیچ گونه اصلاحی روی بردارهای بازنمایی گفتار مستقیم ایجاد نشده بود طبیعی و قابل انتظار است.

۷-٤- اصلاح بردارهای بازنمایی با استفاده از معکوس سازی عمومی

در این روش، هدف این است که معکوس شبکه بازشناسی دو منظوره با یک مدل شبکه عصبی MLP دیگر تخمین زده شود. ساختار این شبکه در شکل ۹ رسم شده است و مبانی نظری مربوط به معکوس‌سازی عمومی نیز در بخش ۶-۲ تبیین شده است. شبکه دارای دو لایه پنهان با توابع غیرخطی تانژانت هایپربولیک و لایه خروجی با نورونهای خطی می‌باشد. سایر پارامترهای شبکه مشابه شبکه مرجع در نظر گرفته شده است. برای تعلیم این شبکه همانگونه که ذکر شد ورودیهای آن را خروجیهای مدل دومنظوره شکل ۸ و اهداف خروجی آن را بردارهای بازنمایی نظیر آنها قرار می‌دهیم.

پس از تعلیم شبکه معکوس، نتایج بازشناسی آواهای‌تلفنیِ شبکه مدل دومنظوره را هم برای دادگان تعلیم و هم برای دادگان آزمون به ورودی شبکه معکوس می‌دهیم و کد دوبیتی مربوط به نوع کانال گفتار را بجای "تلفنی" بودن "مستقیم" قرار می‌دهیم خروجیهای بدست آمده از شبکه معکوس یک مجموعه بردارهای بازنمایی جدید است که می‌توان گفت مجموعه بردارهای بازنمایی مستقیمی است که از روی دادگان تلفنی تخمین زده شده است. به همین ترتیب نتایج بازشناسی گفتار مستقیم در خروجی شبکه مدل دومنظوره را برای هم دادگان تعلیم و هم دادگان آزمون به شبکه معکوس می‌دهیم تا تخمین شبکه معکوس از بردارهای بازنمایی گفتار مستقیم نیز بدست آید.

حال دادگان تعلیم جدید بدست آمده را به شبکه MLP دیگری مشابه شبکه مرجع، تعلیم می‌دهیم و صحت بازشناسی آواهای گفتار را برای دادگان آزمون جدید بررسی می‌نماییم. نتایج آزمون این شبکه جدید در جدول(۳) آمده است. افزایش صحت بازشناسی دادگان تلفنی در این روش نسبت به شبکه مرجع ۲/۸۹٪ و افزایش صحت در بازشناسی دادگان مستقیم برابر ۱/۶۸٪ می‌باشد. به این ترتیب روش معکوس‌سازی عمومی در اصلاح دادگان موفق‌تر از معکوس‌سازی به روش گرادیان بوده است.

جدول ۳: نتایج بازشناسی آواهای گفتار مستقیم و تلفنی بر حسب درصد.

شبکه عصبی	صحت بازشناسی مستقیم	صحت بازشناسی تلفنی	تشخیص مستقیم یا تلفنی
مرجع	۸۴/۱±۰/۲	۶۹/۷۶±۰/۱	---
مدل دو منظوره	۸۳/۹۲	۶۹/۳۹	۹۹/۹۵- ۱۰۰
معکوس گرادیان	۸۳/۹۴	۷۱/۲۴	---
معکوس عمومی	۸۵/۷۸	۷۲/۶۵	---

۸- بحث و تحلیل نتایج

گفتار تلفنی و مستقیم محتوی دو دسته اطلاعات هستند؛ بخش اول، اطلاعات آوایی گفتار است که در هر دو دسته دادگان مشترک است با این تفاوت که در گفتار تلفنی نسبت به گفتار مستقیم قسمتی از این اطلاعات از بین رفته است. بخش دوم، اطلاعات غیرآوایی گفتار یا همان تنوعات گفتار است که ناشی از اثرات گوینده، میکروفون، کانال و غیره است که برای این دو دسته دادگان متفاوت می‌باشد.

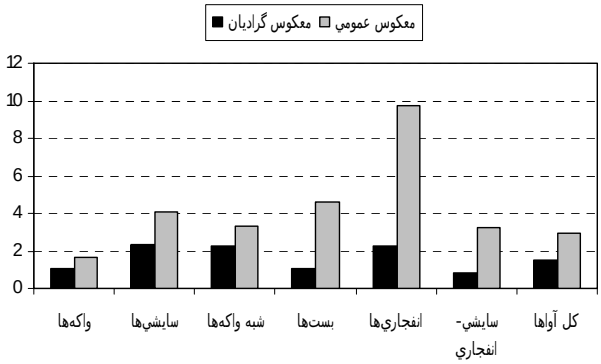
تعلیم یک شبکه عصبی واحد روی این دو دسته دادگان به عنوان مدل مرجع در اولین گام کمک می‌کند تا شبکه به‌عنوان یک واحد محاسباتی موازی و گسترده[1]، بخش مشترک اطلاعات این دو دسته دادگان (اطلاعات آوایی) را بیرون بکشد. و برای گفتار تلفنی از دانش گفتار مستقیم بهره برده، صحت بازشناسی بهتری نسبت به یک سیستم بازشناسی تلفنی به تنهایی داشته باشد (نتایج جدول ۲ برای بازنماییهای LHCB). به این ترتیب در اولین مرحله یک گام به سمت مقاوم‌سازی سیستم بازشناسی گفتار برداشته‌ایم.

مدل بازشناسی دومنظوره از روی بردارهای بازنمایی گفتار که دربردارنده تنوعات مختلف از قبیل گوینده، آواها، لهجه‌ها و نوع کانال است، اطلاعات مشترک آوایی و نوع کانال تلفنی یا مستقیم (غیرمشترک) را شناسایی می‌کند. در معکوس سازی به روش گرادیان با شروع از مقدار اولیه که همان بردار بازنمایی تلفنی است تدریجاً به سمت بردار بازنمایی جدیدی حرکت می‌کند که ضمن حفظ اطلاعات آوایی قبلی‌اش اثر کانال از روی بردارهای بازنمایی تلفنی تا حدی جبران گردد. بنابراین تنها یکی از تنوعات گفتار (نوع کانال) تا حدودی حذف شده است.

در روش معکوس‌سازی عمومی از روی اطلاعات آوایی و نوع کانال گفتار که در خروجی مدل دو منظوره بدست آمده بود، توسط شبکه معکوس، بردارهای بازنمایی دوباره بازسازی می‌شوند. از آنجایی که در ورودی این شبکه غیر از اطلاعات کانال بسیاری از تنوعات گفتار تا حدودی حذف شده‌اند و برای اصلاح بردارهای بازنمایی نیز نوع کانال روی کانال گفتار مستقیم تنظیم می‌شود. بنابراین بردارهای بازنمایی جدیدِ تلفنی و مستقیم بیشتر حاوی اطلاعات آوایی گفتار هستند و اثر کانال تلفنی نیز بصورت خاص از روی بردارهای بازنمایی تلفنی حذف شده است. به این ترتیب در روش معکوس‌سازی عمومی برای هر دو نوع داده اصلاح شده مستقیم و تلفنی، صحت بازشناسی بالاتری نسبت به شبکه مرجع بدست آمده است.

برای نشان دادن موفقیت این دو روش معکوس‌سازی در بهبود بردارهای بازنمایی گفتار تلفنی، شکل ۱۰ میزان افزایش صحت بازشناسی گفتار تلفنی برای شش دسته آوایی مختلف نسبت به شبکه مرجع را

1. Parallel and Distributed Computation



شکل ۱۰: درصد افزایش بازشناسی دسته‌های آوایی مختلف گفتار تلفنی.

نشان می‌دهد و بهترین بهبود در هر دو روش برای آواهای انفجاری گفتار بدست آمده است. آن‌چه که مطابق شکل ۷ یکی از بیشترین دسته‌های آوایی تخریب شده توسط کانال تلفنی است.

نکته مهم و لازم به ذکر این است که اگرچه درصد حضور آواهای مختلف در دادگان تعلیم و آزمون شبکه‌ها متفاوت هستند (به عنوان مثال بیشترین تعداد را واکه‌ها و کمترین تعداد را سایشی-انفجاری‌ها تشکیل می‌دهند) اما از آنجایی که در تمام مراحل این تحقیق، دادگان تعلیم و آزمون شبکه‌ها تغییر نکرده‌اند بلکه همان دادگان از روی دانش آموخته شده در شبکه اصلاح شده و دوباره به مدل جدید تعلیم داده شده‌اند بنابراین تفاوت درصد حضور آواهای مختلف به توفیق این روش‌ها در بهبود بردارهای بازنمایی خدشه‌ای وارد نمی‌کنند.

۹- نتیجه‌گیری

در این تحقیق به منظور طراحی یک سیستم واحد بازشناسی گفتار مستقیم و تلفنی سعی گردید بردارهای بازنمایی گفتار بصورت بهینه انتخاب گردد. نشان داده شد که بردارهای بازنمایی طیفی LHCB برای بازشناسی گفتار تلفنی و بازشناسی توأم گفتار مستقیم و تلفنی در مدلهای بازشناسی مبتنی بر شبکه‌های عصبی مناسب‌تر از بردارهای بازنمایی رایج MFCC هستند. آنگاه با استفاده از معکوس‌سازی شبکه‌های عصبی به روش گرادیان بردارهای بازنمایی گفتار تلفنی به سمت بردارهای بازنمایی گفتار مستقیم اصلاح گردید و با تعلیم شبکه دیگری روی دادگان اصلاح شده تلفنی و دادگان مستقیم دست نخورده، افزایش۱/۴٪ در صحت بازشناسی گفتار تلفنی حاصل گردید. در مرحله بعد با استفاده از معکوس سازی عمومی شبکه‌های عصبی هر دو دسته بردارهای بازنمایی گفتار مستقیم و تلفنی به گونه‌ای اصلاح شدند که بیشتر حاوی اطلاعات آوایی گفتار باشند و سایر تنوعات تا جای ممکن حذف شوند. سپس با تعلیم شبکه دیگری روی این دادگان اصلاح شده، افزایش ۲/۹۸٪ در صحت بازشناسی گفتار تلفنی و ۱/۶۸٪ در صحت بازشناسی گفتار مستقیم بدست آمد.

مراجع

[1] S. Fouri, "Robust methods in automatic speech recognition and understanding," in *Proc. Eurospeech*, pp. 1993-1997, Geneva, Switzerland, 2003.

[2] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Communication*, vol. 16, no. 3, pp. 261-291, Apr. 1995.

[3] C. H. Lee and Q. Huo, "On adaptive decision rules and decision parameter adaptation for automatic speech recognition," in *Proceedings of the IEEE*, vol. 88, pp. 1241- 1269, Aug. 2000.

[4] A. Martin, J. Fiscus, B. Fisher, D. Pallet, and M. Przybocki, "System Descriptions and Performance Summary," presented at the *Conversational Speech Recognition Workshop: DARPA Hub-5E Evaluation*, Baltimore, Maryland, US, May 1997.

[5] D. Yuk and J. Flanagan, "Telephone speech recognition using neural networks and hidden Markov models," in *Proc. ICASSP*, pp. 157-160, 1999.

[6] S. Thrun, "Is learning the n-th thing any easier than learning the first?" *Advances in Neural Information Processing Systems*, MIT Press, 1996.

[7] S. Ben-David and R. Schuller, "Exploiting task relatedness for multiple task learning," *Lecture Notes in Computer Science,* vol. 2777, pp. 567 - 580,2003.

[8] C. W. Omlin and C. L. Giles, "Training second-order recurrent neural networks using hints," in *Proc. of the Ninth International Conference on Machine Learning.*, pp. 363-368, 1992.

[9] S. Parveen and P. Green, "Multitask learning in connectionist robust ASR using recurrent neural networks," in *Proc. Eurospeech*, pp. 1813-1816, Geneva, Switzerland, Sep. 2003.

[10] P., Niyogi and *et al*. "Incorporating prior information in machine learning by creating virtual examples," in *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2196-2209, Nov. 1998.

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

[12] A. Blumer, A. Ehrenfeucht, D. Haussler, and M.Warmuth, "Learnability and the Vapnik-Chervo-Nenkis dimention," *J. Ass. Comput. Match.*, vol.36, no.4, pp. 929-965, 1989.

[13] M. Bijankhan, J. Seikhzadeghan, M. R. Roohani, Y. Samareh, K. Lucas, M. Tebyani., "FARSDAT: the speech database of Farsi spoken language," in *Proc. SST-94*, pp. 826-831, Perth, Australia, 1994.

[14] S. B. Davis and P. Mermelstein, "Comparison of parametric representations of monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. ASSP*, vol. 28, no. 4, pp. 357-366, Aug. 1980.

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

[16] J. Han and W. Gao, "Robust telephone speech recognition based on channel compensation," *Journal of Pattern Recognition Society*, vol. 32, no.6, pp. 1061-1067, Jun. 1999.

[17] C. A. Jensen, *et al*., "Inversion of feedforward neural networks: algorithms and applications," *Proceedings of the IEEE*, vol. 87, no. 9, pp. 1536-1549, Sep. 1999.

[18] R. J., Williams, "Inverting a connectionist network mapping by backpropagation of error," in *Proc 8th Annu. Conf. Cognitive Science Society*, pp. 859-865, 1986.

منصور ولی مدرک کارشناسی خود را در مهندسی برق- الکترونیک از دانشگاه صنعتی اصفهان در سال ۱۳۷۶ و مدرک کارشناسی ارشد را از دانشگاه صنعتی شریف در مهندسی برق- بیوالکتریک در سال ۱۳۷۸ دریافت نموده است. وی در حال حضر دانشجوی دوره دکتری مهندسی پزشکی- بیوالکتریک در دانشگاه صنعتی امیرکبیر می‌باشد. زمینه‌های علمی مورد علاقه ایشان پردازش سیگنال‌های حیاتی، پردازش گفتار و شبکه‌های عصبی مصنوعی و زیستی می‌باشد.

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