بازشناسی مقاوم گفتار تلفنی با ترکیب روش اصلاح بردارهای بازنمایی توسط شبکه عصبی دو سویه و روشهای MLLR و MAP در تطبیق مدل بازشناسی

ایمان اسمعیلی [*1]، منصور ولی [2]
[2،1]دانشگاه شاهد
[1]i.esmaili@yahoo.com, [2]vali@shahed.ac.ir

چکیده - سیستمهای تعلیم یافته با گفتار میکروفنی باند پهن در محیط تلفنی دارای صحت بازشناسی بسیار پایینی هستند. در این مقاله ابتدا با استفاده از شبکه عصبی دو سویه بردارهای بازنمایی گفتار تلفنی در جهت انطباق با بردارهای بازنمایی گفتار میکروفنی اصلاح می شوند. در مرحله بعد مدل جدیدی با استفاده از بردارهای اصلاح شده آموزش داده می شود و بردارهای بازنمایی یک مرحله دیگر اصلاح می شوند. بردارهای بازنمایی یک مرحله اصلاح شده و دو مرحله اصلاح شده برای تعلیم مدلهای مخفی مارکف به کار گرفته شده اند. در این مرحله درصد بازشناسی مدلهای تعلیم یافته توسط بردارهای بازنمایی میکروفنی که توسط بردارهای بازنمایی یک مرحله اصلاح شده و دو مرحله اصلاح شده تلفنی ارزیابی شده اند به ترتیب 22/3 درصد و 26/6 درصد افزایش می یابد. در ادامه کار روش اصلاح بردارهای بازنمایی با تکنیک شبکه عصبی دوسویه با روشهای معمول تطبیق مدل (MLLR، MAP وMAP + MLLR) ترکیب می شود. درصد بازشناسی ترکیب بردارهای اصلاح شده با تکنیکهای MLLR، MAP و MAP + MLLR به ترتیب 37/7 درصد، 39/6 درصد و 40/2 درصد نسبت به مدل آموزش دیده با بردارهای بازنمایی اصلاح نشده افزایش می یابد.

کلید واژه- اصلاح بردارهای بازنمایی، تطبیق مدل، شبکه عصبی دوسویه، مدل مخفی مارکف.

1- مقدمه

عدم انطباق داده های آموزش و آزمون یکی از دلایل اصلی کاهش بازدهی سیستمهای بازشناسی گفتار است. در میان انواع علل ممکن برای ایجاد این عدم انطباق، در این مقاله به روشهای مقابله با عوامل ایجاد عدم انطباق در باند گفتار تلفنی پرداخته می شود.

با فرض وجود داده های کافی از محیط آزمون سیستم بازشناسی گفتار، ساده ترین راه حل غلبه بر عدم انطباق، آموزش دوباره مدل با داده های محیط آزمون است. مهمترین محدودیت این روش فراهم آوردن داده های آموزش کافی از محیط جدید و همچنین زمان بر بودن تعلیم مجدد سیستم بازشناسی گفتار است. هنگامیکه عوامل تخریب بخشهای مختلفی از سیگنال صوتی را تحت تاثیر قرار می‌دهند روشهای مبتنی بر تطبیق مدل نیازمند اصلاح اساسی در پارامترهای مدل هستند و هر چه میزان عوامل تخریب متنوع تر باشد، تطبیق مدل زمان و حجم محاسبات بیشتری را می طلبد. در اینگونه موارد استفاده از روشهای تطبیق بردارهای بازنمایی روش مناسب تری است.

روش‌های تطبیق بردار های بازنمایی، عموماً در بلوکی مجزا از سیستم بازشناسی، بردارهای بازنمایی محیط آزمون را با محیط آموزش تطبیق می‌دهند. این روشها به طور گسترده در محیطهای نویزی مورد استفاده قرار گرفته اند و اخیراً در مورد سیگنالهای باند محدود نیز به کار گرفته شده اند

یکی دیگر از تکنیک های موثری که در سالهای اخیر برای بازشناسی سیگنالهای آغشته به نویز معرفی شده است تکنیک داده های مفقود است [1]. این روش شامل دو مرحله است در مرحله اول بخش-های تخریب شده و سالم طیف گفتار شناسایی می شوند و در مرحله دوم قسمت‌های تخریب شده طیف یا کنار گذاشته می شوند و یا از روی قسمت‌های سالم بازسازی می شوند. آنچه باعث می شود که نتوان این تکنیک را در مورد گفتار تلفنی به کار گرفت این واقعیت است که هیچگونه فریم تخریب نشده ای در باندهای مفقود شده گفتار تلفنی وجود ندارد که به وسیله آن بتوان فریمهای تخریب شده را بازسازی کرد. ما در کار قبلی خود [2] برای تحقق این امر دادگان گفتار میکروفنی را در کنار دادگان گفتار تلفنی بکار گرفتیم. به این ترتیب که بردارهای بازنمایی استخراج شده ازگفتار میکروفنی و تلفنی(لگاریتم

انرژی بانکهای فیلتر)، بصورت توأمان به یک مدل شبکه عصبی دوسویه تعلیم داده می شود و در جهت افزایش صحت بازشناسی اصلاح می شوند. بخشهای مفقود طیفی در بردارهای بازنمایی تلفنی نیز از روی دانش نهفته در شبکه ناشی از یادگیری توأمان بردارهای بازنمایی گفتار میکروفنی و تلفنی تخمین زده می‌شوند.

در مقاله [3] برای جبران بردارهای بازنمایی گفتار باند محدود، ابتدا بردارهای بازنمایی گفتار باند محدود خوشه بندی می شود و توابع انتقال برای هر خوشه به طوری که بردارهای بازنمایی باند محدود به بردارهای بازنمایی باند پهن منتقل شود، تخمین زده می شوند. هر چند این روش برای گفتار باند محدودی که به طور مصنوعی ساخته شده است درصد بازشناسی قابل قبولی دارد اما در مورد سیگنالهای گفتار تلفنی واقعی نه تنها بهبودی نشان نمی دهد بلکه نتایج بازشناسی پایین تری هم بدست می آید. اما در روش ارائه شده در این مقاله شبکه عصبی دوسویه با گفتار تلفنی و میکروفونی به صورت همزمان تعلیم داده می شود و در نتیجه با استفاده از دانش آوایی مشترک موجود در گفتار تلفنی و میکروفونی به خوبی می توان بردارهای بازنمایی را اصلاح نمود.

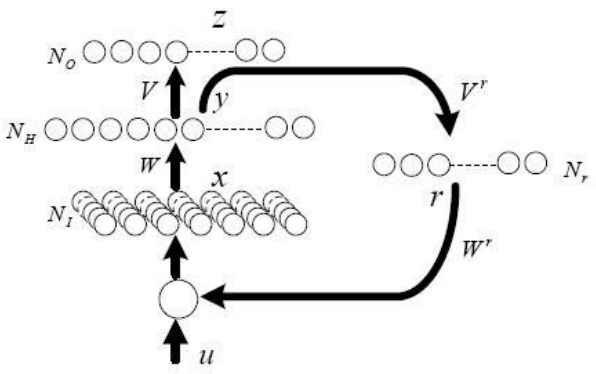
ما در این مقاله به منظور جبران هر چه بیشتر بردارهای بازنمایی گفتار تلفنی و میکروفونی ، بردارهای بازنمایی اصلاح شده توسط شبکه عصبی دوسویه را یکبار دیگر با همان روش اصلاح کرده و بردارهای دو مرحله اصلاح شده را بدست آورده ایم. ما همچنین بردارهای بازنمایی MFCC را با اعمال تبدیل کسینوسی گسسته بر روی بردارهای بازنمایی لگاریتم انرژی بانکهای فیلتر بدست آورده ایم و این بردارهای بازنمایی را برای تعلیم مدلهای مارکف به کار گرفته ایم.

از جمله متداول ترین روشهای تطبیق مدل که به همراه مدل مخفی مارکف مورد استفاده قرار می گیرند روشهای رگرسیون خطی بیشینه شباهت (MLLR) [4] و بیشینه احتمال پسین (MAP) [5] هستند. در این مقاله ابتدا روش اصلاح بردارهای بازنمایی با استفاده از شبکه عصبی دو سویه با روشهای متداول تطبیق مدل (تعلیم مجدد مدل، MAP، MLLR و MLLR+MAP) مقایسه می شود و در گام بعد روش اصلاح بردارهای بازنمایی و تطبیق مدل با یکدیگر ترکیب می شوند.

نتایج بدست آمده از اصلاح بردارهای بازنمایی با تکنیک شبکه عصبی به خوبی توانایی این تکنیک را در بازسازی بردارهای بازنمایی گفتار تلفنی به نمایش می گذارد. نتایج حاصل از ترکیب تکنیک شبکه عصبی دوسویه با روش های تطبیق مدل از نتایج تعلیم مجدد مدل با داده های محیط آزمون هم فراتر می رود و این موضوع نشانگر این امر است که فرضیه ما در مورد استفاده شبکه عصبی از دانش نهفته در بردارهای بازنمایی میکروفنی برای اصلاح بردارهای بازنمایی تلفنی صحیح است.

2- شبکه عصبی دو سویه

ساختار شبکه دوسویه مطابق شکل 1 مشتمل بر یک شبکه یک سویه چند لایه ، به همراه یک شاخه بازگشتی از لایه مخفی آن به ورودی است. این شاخه بازگشتی خود شامل یک لایه مخفی با N_r نورون، تابع غیرخطی نوع تانژانت هایپربولیک و وزن‌های تمام متصل W^r و V^r می‌باشد. خروجی شاخه بازگشتی خود یک بردار N_I مؤلفه‌ای است که با مؤلفه‌های نظیر خود از بردارهای بازنمایی ورودی ترکیب می‌شوند.



شکل 1: ساختار شبکه عصبی دو سویه

هدف از بکارگیری مسیر بازگشتی در شبکه عصبی دوسویه این است که از روی دانش یادگیری شده در لایه مخفی بخش جلوسوی این شبکه مؤلفه‌های بازنمایی تخریب شده یا مفقود شده در ورودی اصلاح شوند. دلیل فراهم شدن چنین امکانی در این شبکه برمی‌گردد به این نکته که شبکه جلوسو برای هر آوای گفتار ویژگیهای سالم از گفتار میکروفنی را در اختیار دارد که به علت همبستگی بهتر آنها و تعداد بیشتر این ویژگیها نسبت به ویژگیهای تخریب شده و پراکنده تلفنی، آنها را بهتر یاد می گیرد. علاوه براین، دانش یادگرفته شده در لایه مخفی شبکه بیشتر حاوی اطلاعات آوایی گفتار است و بسیاری از پراکندگی‌های مابین ویژگیهای ورودی که برای بازشناسی آواها مفید نمی‌باشند، در آن سطح از بین رفته است. بنابراین برای بهبود و اصلاح بردارهای بازنمایی ورودی بهترین اطلاعات در این سطح شبکه قرار گرفته است. در مسیر بازگشتی شبکه عصبی دوسویه از روی این دانش، مؤلفه‌های اصلاحی برای هر یک از مؤلفه‌های بردارهای بازنمایی ورودی بوسیله یک تابع غیرخطی تخمین زده می‌شوند و با یکدیگر جمع می‌شوند.

الگوریتم تعلیم شبکه عصبی دوسویه به طور مفصل در [2] بحث شده است. بنابراین در اینجا تنها خلاصه ای از نحوه عملکرد آن توضیح

داده می شود. روش کلی تعلیم شبکه دوسویه شکل 1 مشابه شبکه عصبی جلوسوی مرجع است. یعنی الگوریتم تعلیم همان پس انتشار خطا در جهت کاهش گرادیان خطاست. به این ترتیب وزنهای جلوسو و بازگشتی شبکه به منظور بازشناسی محتوای آوایی الگوهای ورودی اصلاح می‌شوند. تفاوت الگوریتم تعلیم نسبت به یک شبکه عصبی چند لایه معمولی در این است که ورودیهای این شبکه در هر دور تعلیم برابر حاصلجمع ورودی فعلی و مقادیر اصلاحی محاسبه شده از روی مقادیر قبلی لایه مخفی شبکه هستند. در هر دور تعلیم کلیه فریمهای دادگان تعلیم وارد شبکه شده و وزنهای شبکه، مطابق روابط الگوریتم تعلیم طبق رابطه (1) اصلاح می شوند.

$$x_i[n] = \lambda u_i + \sum_{l=0}^{N_r} r_l[n] w_{li}^r \qquad i = 1,2,..., N_I \quad (۱)$$

رابطه (1) نشان می‌دهد که حاصلجمع کسری از بردار بازنمایی (λu در حالیکه $0< \lambda <1$) و ترکیب خطی مقادیر لایه مخفی بازگشتی شبکه (r)، ورودی شبکه را در دور n ام تعلیم تشکیل می‌دهد. در ابتدای تعلیم (n =1) مقدار ورودی شبکه $x_i[1] = u$ در نظر گرفته می‌شود. لازم به ذکر است که اگرچه این مؤلفه‌های اصلاحی با مؤلفه‌های بردارهای بازنمایی ورودی جمع می‌شوند اما به دلیل اینکه ماهیت ویژگیها که لگاریتم انرژی طیف گفتار در باندهای فرکانسی مختلف هستند، عبارت جمعی فوق معادل ضرب هر یک از انرژیهای باندهای فرکانسی در یک ترم جبرانساز ناشی از تخمین مسیر بازگشتی شبکه است.

3- تطبیق مدل

یکی از روشهای پر کاربرد بازشناسی مقاوم گفتار، تطبیق پارامترهای مدل بازشناسی مرجع با محیط آزمون است. زمانیکه از مدل مخفی مارکف در سیستم بازشناسی گفتار استفاده شود، تکنیکهای MLLR و MAP و یا ترکیب آنها می توانند به طور موثری برای تطبیق محیطهای آموزش و آزمون مورد استفاده قرار گیرند.

3-1- تطبیق مدل با MLLR

رگرسیون خطی بیشینه شباهت (MLLR) با محاسبه مجموعه ای از تبدیلات خطی، موجب کاهش عدم انطباق بین پارامترهای مدل اولیه و داده های آزمون می شود. در دستگاههای بازشناسی مبتی بر مدل مخفی مارکف، اعمال این تبدیلات خطی به نحوی میانگین ها و واریانس های مدل اولیه را تغییر می دهد که حالتهای مدل مخفی مارکف با احتمال بیشتری داده های آموزش شبکه را تولید کنند.

ماتریس تبدیل برای تخمین میانگین های جدید مدل از فرمول (2) محاسبه می شود:

$$\hat{\mu} = W\xi \qquad (2)$$

در این معادله W ماتریس تبدیل $n \times (n+1)$ است که n تعداد مولفه های بردار داده های آموزش مدل است و ξ مطلبق معادله (3) برداری است که مولفه اول آن یک مقدار ثابت بایاس و بقیه مولفه ها مقادیر میانگین هستند.

$$\xi = \begin{bmatrix} b & \mu_1 & \mu_2 & \dots & \mu_n \end{bmatrix} \qquad (3)$$

اگر داده های مورد استفاده برای انطباق مدل کم باشد تنها یک تبدیل کلی برای تمام میانگین های و برای همه توابع چگالی گوسی در نظر گرفته می شود. با بالا رفتن میزان داده‌های تطبیق مدل می توان تبدیلات بیشتری در نظر گرفت. یکی از راههای موثر انتخاب میزان تبدیلات استفاده از درخت تصمیم گیری است. در این روش توابع چگالی گوسی بسته به نوع و میزان داده های مورد استفاده در تطبیق مدل گروه بندی می شوند و برای مجموعه تبدیلات کلاس های مختلف محاسبه می شود. استفاده از این روش موجب می شود تا پارامترهایی که هیچگونه داده آموزشی برای آنها موجود نیست هم با داده های آموزشی انطباق یابند.

3-2- تطبیق با MAP

در روش MAP از اطلاعات پیشین توزیع پارامترهای مدل برای تطبیق با داده‌های آموزشی استفاده می شود. معادله (4) نحوه تطبیق تابع چگالی گوسی m ام از حالت j ام مدل مخفی مارکف را نشان می دهد:

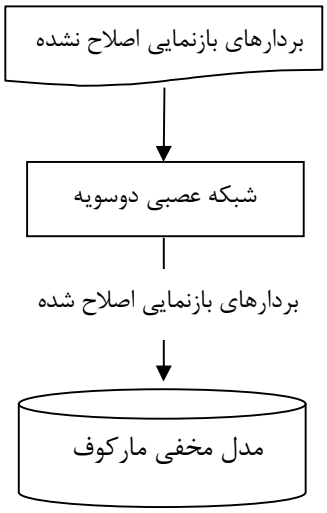
$$\hat{\mu} = \frac{N_{jm}}{N_{jm} + \tau} \bar{\mu}_{jm} + \frac{\tau}{N_{jm} + \tau} \mu_{jm} \qquad (4)$$

در این معادله τ برای وزن دهی دانش پیشین نسبت به داده های انطباق مدل و N احتمال وقوع داده های انطباق است.

روشهای زیادی برای ترکیب MAP و MLLR وجود دارد. در این مقاله این دو روش با در نظر گرفتن میانگین تبدیلات MLLR به عنوان دانش پیشین برای روش تطبیق MAP (با جایگزین کردن μ_{jm} از معادله (4) با میانگین تبدیلات معادله (2)) با یکدیگر ترکیب شده اند.

4- سیستم بازشناسی گفتار

اجزاء اصلی سیستم بازشناسی گفتار ارائه شده در این مقاله در شکل 2 نشان داده شده است.



شکل 2: بلوک دیاگرام سیستم بازشناسی گفتار

پس از تعلیم شبکه عصبی بردارهای آزمون به ورودی شبکه اعمال می شوند. این بردارها برای N، . . ., 2، 1 = n به شبکه اعمال می شوند که n تعداد دورهایی است که دادگان در شبکه به گردش در می آیند. آزمایشات نشان داده است که سه دور گردش در مدل برای اصلاح بردارهای بازنمایی کفایت می کند [2]. در گام بعد این بردارها برای بازشناسی، به مدلهای مخفی مارکوف که با بردارهای اصلاح شده تعلیم داده شده اند، اعمال می شوند.

5- پیاده سازی مدل بازشناسی

5-1- دادگان گفتار

دادگان گفتاری مورد نیاز برای تعلیم شبکه‌ها از مجموعه دادگان فارس‌دات [6] و فارس‌دات تلفنی [7] انتخاب شده‌اند. این مجموعه دادگان شامل 128جمله بیان شده توسط 64 گویشور تلفنی و 400 جمله بیان شده توسط 200 گویشور مستقیم (میکروفونی) می‌باشد. نرخ نمونه برداری سیگنالهای گفتار برای گویشهای تلفنی 8 کیلوهرتز و برای گویشهای میکروفونی برابر 16 کیلوهرتز است. هفتاد و پنج درصد از مجموعه دادگان برای آموزش و بیست و پنج درصد باقیمانده برای آزمون سیستم بازشناسی در نظر گرفته شده است.

5-2- استخراج بردارهای بازنمایی

لگاریتم انرژی بانکهای فیلتر (LFBE) به عنوان برادارهای بازنمایی در نظر گرفته شده اند. برای استخراج این بردارهای بازنمایی یک بانک فیلتر 18 تایی در فاصله صفر تا 8 کیلو هرتز در نظر گرفته شده است. محدوده فرکانسی مناسب برای استخراج بردارهای بازنمایی از گفتار تلفنی (دارای محدوده فرکانسی بین صفر تا 4 کیلوهرتز) را معمولا بین 125 تا 3700 هرتز انتخاب می‌کنند. 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. بنابراین از 18 بانک فیلتر طراحی شده برای گفتار میکروفنی بانکهای فیلتر دوم تا چهاردهم انتخاب شده و با محاسبه لگاریتم انرژی این 13 بانک فیلتر، بردار بازنمایی 13 تایی تلفنی بدست می‌آید. با اضافه شدن مشتق اول و دوم طول بردارهای بازنمایی به 39 و 54 به ترتیب برای گفتار تلفنی و میکروفنی افزایش می یابد. از آنجا که برای اعمال بردارهای بازنمایی به شبکه عصبی دوسویه طول بردارها باید یکسان باشد به فیلترهای مفقود شده تلفنی (فیلتر اول و فیلترهای پانزدهم تا هجدهم) مقدار صفر تخصیص داده می شود. این مقادیر مفقود شده در فرایند اصلاح بردارهای بازنمایی توسط شبکه عصبی دو سویه با مقدار مناسب، بارگذاری خواهند شد.

5-3- شبکه عصبی دوسویه

یک شبکه عصبی دوسویه مطابق شکل 1 برای جبران بردارهای بازنمایی در نظر گرفته شده است. ورودی شبکه عصبی شامل 7 فریم متوالی گفتار (فریم جاری به علاوه سه فریم قبل و سه فریم بعدی آن) می باشد به عبارت دیگر 7 فریم با بردارهای بازنمایی 54 تایی که 378 نورون ورودی خواهد شد. لایه مخفی شبکه مشتمل بر 100 واحد و خروجی شبکه شامل 34 نورون متناظر با 34 آوای در نظر گرفته شده گفتار فارسی است. تعداد نورون‌های بخش بازگشتی شبکه عصبی نیز برابر با 54 واحد در نظر گرفته شده است.

5-4- مدلهای مخفی مارکف

مدل مخفی مارکوف برابر با 34 آوی فارسی، برای هر دسته از بردارهای بازنمایی آموزش داده شده است. مدلهای مخفی مارکوف شامل سه حالت چپ به راست به همراه شانزده تابع چگالی گوسی برای هر حالت هستند.

5-5- تطبیق مدل

چهار روش روش رایج تطبیق مدل(تعلیم مجدد مدل، MLLR، MAP و MLLR+MAP) در این مقاله مورد استفاده قرار گرفته اند. در تطبیق مدل به روش تعلیم مجدد مدل، از 96 جمله فارس دات تلفنی استفاده شده است. در تطبیق به روش MLLR در آزمایشهای مختلف از 10 تا 96 جمله جهت تطبیق مدل استفاده شده است. در این مقاله برای حصول به نتایج بهتر MLLR در دو مرحله انجام شده است

درمرحله اول، MLLR با یک کلاس داده و در مرحله دوم MLLR با 32 کلاس داده بر مدلهای مرحله اول اعمال شده است. در تطبیق به روش MAP نیز در آزمایشهای مختلف از 10 تا 96 جمله برای تطبیق مدلها استفاده شده است. همچنین روش MAP بر پارامترهای اصلاح شده توسط MLLR اعمال شده اند و نتایج تطبیق پارامترهای مدل با ترکیب این دو روش نیز مورد بررسی قرار گرفته اند.

5-6- ارزیابی

مطابق شکل 2، بردارهای بازنمایی بوسیله شبکه عصبی دوسویه اصلاح شده اند. سپس این بردارهای اصلاح شده، برای تعلیم مدلهای مارکف مورد استفاده قرار گرفته اند. به منظور محک زدن حداکثر قدرت شبکه عصبی در جبران بردارهای بازنمایی، شبکه عصبی دوسویه جدیدی با بردارهای اصلاح شده تعلیم داده می شود و با اعمال بردارهای اصلاح شده به ورودی آن، بردارهای بازنمایی یک مرحله دیگر نیز اصلاح می شوند. بردارهای بازنمایی MFCC را نیز با اعمال تبدیل کسینوسی گسسته بر بردارهای بازنمایی لگاریتم انرژی بانکهای فیلتر بدست آمد ه اند.

به منظور ارزیابی سیستم بازشناسی پیشنهادی در شرایط عدم انطباق داده های آزمون و آموزش مدلهای زیر در نظر گرفته شده اند:

مدل A: آموزش توسط داده های اصلاح نشده میکروفونی و آزمون توسط داده های اصلاح نشده تلفنی.

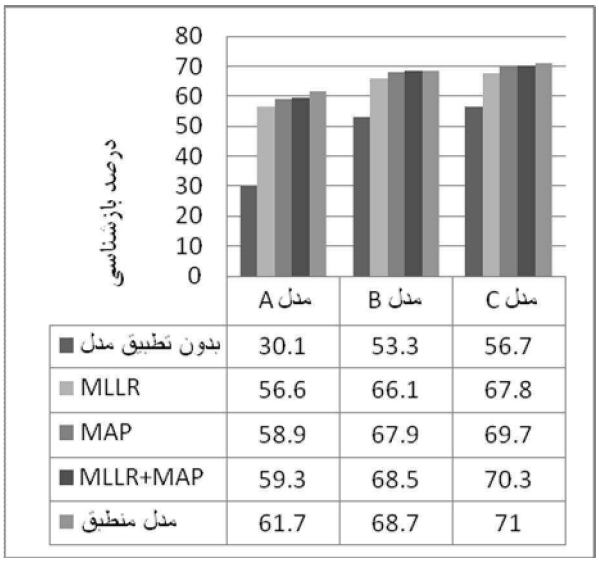
مدل B : آموزش توسط داده های یک مرحله اصلاح شده میکروفونی و آزمون توسط داده های یک مرحله اصلاح شده تلفنی.

مدل C : آموزش توسط داده های دو مرحله اصلاح شده میکروفونی و آزمون توسط داده های دو مرحله اصلاح شده تلفنی.

مدل منطبق: آموزش و تست توسط بردارهای بازنمایی تلفنی(تطبیق مدل به روش آموزش مجدد)

درصد بازشناسی آوا برای مدلهای A، B و C و ترکیب آنها با روشهای مختلف تطبیق مدل با حداکثر داده های تطبیق (96 جمله) در شکل 3 نشان داده شده است. تفاوت زیاد بازشناسی آوای مدل A (30/1 درصد) که بدون اصلاح بردارهای بازنمایی است و مدل منطبق (61/7 درصد) نشان دهنده نیاز ضروری تطبیق شرایط (بردارهای بازنمایی و مدل بازشناسی) محیط آزمون با شرایط محیط آموزش سیستم بازشناسی است. استفاده از بردارهای یک مرحله اصلاح شده و دو مرحله اصلاح شده درصد بازشناسی آوا را به ترتیب 22/3 درصد و 26/6 درصد نسبت به حالت بدون اصلاح مدل افزایش داده است. این امر خود نشانگر توانایی شبکه عصبی دوسویه در اصلاح بردارهای بازنمایی با استفاده از دانش مشترک داده های میکروفنی و تلفنی است.

ترکیب اصلاح بردارهای بازنمایی یک مرحله ای به وسیله شبکه عصبی دوسویه و روشهای MLLR، MAP و MLLR + MAP درصد بازشناسی را نسبت به مدل بدون اصلاح بردار بازنمایی و بدون تطبیق مدل به ترتیب 36 درصد، 37/8 درصد و 38/4 افزایش داده است و افزایش درصد بازشناسی برای ترکیب بردارهای بازنمایی یک مرحله اصلاح شده و روشهای MLLR، MAP و MLLR + MAP نسبت به روشهای متناظر تطبیق مدل بدون اصلاح بردارهای بازنمایی به ترتیب 9/5 درصد، 9 درصد و 9/2 درصد می باشد.



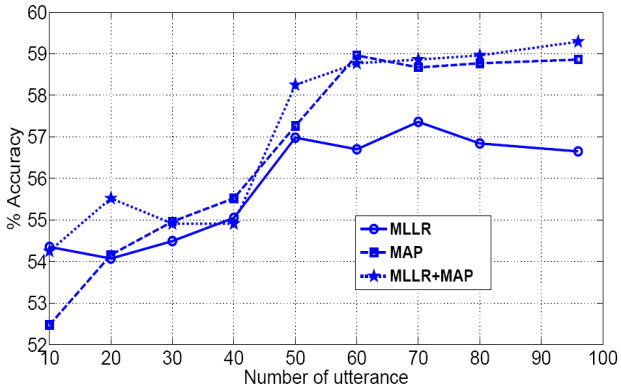
	مدل A	مدل B	مدل C
بدون تطبیق مدل	30.1	53.3	56.7
MLLR	56.6	66.1	67.8
MAP	58.9	67.9	69.7
MLLR+MAP	59.3	68.5	70.3
مدل منطبق	61.7	68.7	71

شکل 3: درصد بازشناسی آوای مدلهای A، B و C و همچنین تطبیق و بهبود این مدل ها با روشهای MLLR، MAP و ترکیب MLLR و MAP به همراه تطبیق مدل به روش تعلیم مجدد (مدل منطبق).

ترکیب بردارهای بازنمایی دو مرحله اصلاح شده به وسیله شبکه عصبی دوسویه و روشهای MLLR، MAP و MLLR + MAP درصد بازشناسی را نسبت به مدل بدون اصلاح بردار بازنمایی و بدون تطبیق مدل به ترتیب 37/7 درصد، 39/6 درصد و 40/2 افزایش داده است و افزایش درصد بازشناسی برای ترکیب بردارهای بازنمایی دو مرحله اصلاح شده و روشهای MLLR، MAP و MLLR + MAP نسبت به روشهای متناظر تطبیق مدل بدون اصلاح بردارهای بازنمایی به ترتیب 11/2 درصد، 10/8 درصد و 11 درصد می باشد.

تاثیر اصلاح بردارهای بازنمایی در مدل منطبق نیز به خوبی مشهود است به طوری که درصد بازشناسی آوا برای بردارهای بازنمایی یک مرحله اصلاح شده و دو مرحله اصلاح شده نسبت به مدل آموزش دیده بوسیله بردارهای بازنمایی اصلاح نشده به ترتیب 7 درصد و 9/3 درصد افزایش یافته است.

در شکل 4، به منظور مقایسه اصلاح بردارهای بازنمایی به روش شبکه عصبی دو سویه و روش تطبیق مدل، درصد بازشناسی آوا بر حسب تعداد جملات تطبیق برای روشهای MLLR، MAP و ترکیب این دو روش بدون اصلاح بردارهای بازنمایی ارائه شده است.



شکل 4: درصد بازشناسی آوا برای روشهای مختلف تطبیق مدل (MLLR، MAP و MLLR+MAP) بر حسب جملات استفاده شده برای تطبیق مدل.

همانگونه که در شکل 4 نشان داده شده است برای تعداد جملات کمتر از 50، بردارهای دو مرحله اصلاح شده به روش شبکه عصبی (56/7 درصد) بهتر از روش تطبیق مدل عمل کرده است. اما با بالا رفتن تعداد جملات به دلیل فراهم آمدن داده های کافی برای تخمین ماتریس های تبدیل، روش تطبیق مدل بهتر عمل می کند. با این وجود ترکیب این دو روش همانگونه که نشان داده شد نتایج بازشناسی را به نحو موثری افزایش می دهد به گونه ای که نتایج حتی از مدل منطبق هم فراتر می رود.

6- نتیجه گیری

سیستمهای بازشناسی گفتار تعلیم داده شده با بردارهای بازنمایی میکروفنی در صورت آزموده شدن با بردارهای بازنمایی تلفنی دارای درصد بازشناسی پایینی هستند. در این مقاله روشی بر مبنای شبکه عصبی دوسویه برای جبران بردارهای بازنمایی تلفنی ارائه شد که به نحو موثری بردارهای بازنمایی تلفنی را برای انطباق با بردارهای بازنمایی میکروفنی اصلاح می کند. در این تحقیق همچنین بردارهای بازنمایی جدیدی با آموزش شبکه عصبی دوسویه توسط بردارهای بازنمایی اصلاح شده و درنظر گرفتن بردارهای بازنمایی اصلاح شده به عنوان ورودی شبکه عصبی دوسویه بدست آمد. بردارهای بازنمایی MFCC هم با گرفتن تبدیل کسینوسی گسسته از بردارهای بازنمایی لگاریتم انرژی بانکهای فیلتر بدست آمدند. با استفاده از بردارهای یک مرحله اصلاح شده و دو مرحله اصلاح شده درصد بازشناسی مدل آموزش دیده با گفتار میکرفنی و آزمون شده با بردارهای بازنمایی تلفنی به ترتیب 22/3 درصد و 26/6 درصد نسبت به مدل با بردارهای بازنمایی اصلاح نشده افزایش یافت.

در ادامه کار روش اصلاح بردارهای بازنمایی به روش شبکه عصبی دوسویه با روشهای تطبیق مدل (MLLR، MAP و ترکیب این روشها) ترکیب شد. درصد بازشناسی آوا در این مرحله برای روشهای MLLR، MAP و ترکیب این دو روش به ترتیب 36 درصد، 37/8 درصد و 38/4 با بردارهای بازنمایی یک مرحله اصلاح شده و 37/7 درصد، 39/6 درصد و 40/2 با بردارهای بازنمایی دو مرحله اصلاح شده نسبت به مدل بدون تطبیق مدل و بدون اصلاح بردارهای بازنمایی، افزایش داشته است. درصد بازشناسی آوا در مدل ترکیبی از مدل آموزش داده شده و آزمون شده با بردارهای بازنمایی تلفنی هم فراتر می رود که این امر فرضیه ما در مورد توانایی شبکه عصبی دوسویه در اصلاح بردارهای بازنمایی تلفنی با استفاده از دانش حاصله از تعلیم توامان مدل با داده های تلفنی و میکروفنی را اثبات می کند.

سپاسگزاری

این تحقیق توسط مرکز تحقیقات مخابرات ایران (ITRC) طبق قرارداد 500/3133/ ت پشتیبانی شده است.

مراجع

[1] B. Raj, M. L. Seltzer, R. M. Stern, "Reconstruction of Missing Features for Robust Speech Recognition," *Speech Communication.*, vol. 43, no. 4, pp. 275–296, 2004.

[2] M. Vali. , S.A. Seyyed Salehi. K. Karimi, "Robust Speech Recognition by Modifying Clean and Telephone Feature Vectors Using Bidirectional Neural Network," *Proc. InterSpeech*, pp. 2072-2075, USA, Sep 2006.

[3] N. Morales, D.T. Toledano, J.H.L. Hansen, J. Garrido, "Feature Compensation Techniques for ASR on Band-Limited Speech," *IEEE Trans. Speech and Audio Process*, vol. 17, no. 4, May 2009.

[4] C.J. Leggetter and P.C. Woodland." Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models", *JOUTnul of Computer Speech and Language,* 9(2):171-185, April 1995.

[5] Jean-Luc Gauvain and C.H. Lee." Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains", *IEEE Transactions on Speech and Audio Processing,* 2(2):291-298, April 1994.

[6] M. Bijankhan, J. Sheikhzadegan, M.R. Roohani, Y. Samareh, C. Lucas, M. Tebyani, "FarsDat – The Speech Database of Farsi Spoken Language," *Proc. 5th Australian Int. Conf. Speech Science and Technology (SST)*, pp. 826-831, Perth, Australia, 1994.

[7] M. Bijankhan, J. Sheykhzadegan, M.R. Roohani, R. Zarrintare, S.Z. Ghasemi, M.E. Ghasedi, "TFarsDat – The Telephone Farsi Speech Database," *Proc. EuroSpeech*, Geneva, Switzerland, 2003.