بازشناسی مقاوم گفتار تلفنی فارسی به روش سری های تیلور برداری *(VTS)*

محسن قدیانی*[۱] ، منصور ولی[۲] ، سارا پورمحمدی[۳]

[۱]دانشجوی کارشناسی ارشد دانشگاه شاهد، [۲]استادیار دانشگاه شاهد، [۳]دانشجوی کارشناسی ارشد دانشگاه شاهد
[1]mohsenghadyani@ymail.com , [2]vali@shahed.ac.ir , [3]pourmohammadi@shahed.ac.ir

چکیده - تکنیک سری های تیلور برداری *(VTS)* از جمله کاراترین روش‌های بازشناسی مقاوم گفتار به شمار می رود که در هر دو حوزه اصلاح بردارهای بازنمایی و اصلاح مدل صوتی بازشناسی کاربرد دارد. مقاله حاضر به شرح این تکنیک برای جبران سازی اثر عوامل مزاحم محیطی از روی بردارهای بازنمایی سیگنال گفتار فارسی میکروفنی و تلفنی، و در نتیجه افزایش نرخ بازشناسی آن ها پرداخته است. به این ترتیب که پس از استخراج بردارهای بازنمایی از گفتار تخریب شده توسط نویز جمعی یا کانال تلفن به روش متداول *LFBE*، این بردارها به کمک تکنیک *VTS* اصلاح شده‌اند. برای بازشناسی، یک مدل مبتنی بر شبکه عصبی *MLP* با دینامیک زمانی توسط دادگان تمیز موجود تعلیم داده شده است. نتایج حاصل از تست این شبکه بر روی بردارهای بازنمایی میکروفنی و تلفنی نشان داده است که استفاده از الگوریتم *VTS* در جبران‌سازی بردارهای بازنمایی، نه تنها منجر به بهبود چشمگیری در بازشناسی گفتار میکروفنی برای *SNR* های پایین خواهد شد (۱۵٪ بهبود برای *SNR=0*)، بلکه نرخ بازشناسی گفتار تلفنی در حالت انطباق بین دادگان تعلیم و آزمون نیز بیانگر پیشرفت مناسبی نسبت به گفتار تلفنی اصلاح نشده خواهد بود.

کلید واژه- بازشناسی مقاوم گفتار، تابع تنوعات محیطی، سری های تیلور برداری، گفتار تلفنی، نویز

۱- مقدمه

هنگامی که از سیستم بازشناسی گفتار آموزش یافته در شرایط آزمایشگاهی در محیط واقعی استفاده شود، راندمان سیستم و نرخ بازشناسی به دلیل عدم انطباق دادگان تعلیمی و داده جمع آوری شده در محیط واقعی به مقدار زیادی کاهش می‌یابد. از این رو مبحث مقاوم‌سازی در برابر تنوعات محیطی به عنوان یکی از ضرورت های کاربردی از زمینه‌های مطرح تحقیقاتی در سال های اخیر بوده است[۱]. بر اساس تقسیم بندی یک سیستم بازشناسی الگو به دو بخش استخراج ویژگی و مدل بازشناسی، مجموعه روش های بازشناسی مقاوم گفتار به دو دسته کلی تقسیم می شوند:

الف- تکنیک‌های مبتنی بر اصلاح بردارهای بازنمایی صوتی *(Feature-Based Techniques)* که هدف از آن‌ها اصلاح و مقاوم سازی بردارهای بازنمایی است.

ب- تکنیک های مبتنی بر اصلاح مدل صوتی بازشناسی *(Model-Based Techniques)* که برای جبران سازی مدل در برابر عوامل مزاحم محیطی کاربرد دارند[۲].

در سال های اخیر استفاده از تکنیک سری های تیلور برداری *(Vector Taylor Series)* جهت مقاوم‌سازی سیستم‌های بازشناسی گفتار در برابر تغییرات محیطی به صورت گسترده‌ای رواج یافته است. ایده سری های تیلور برداری روشی تحلیلی و مبتنی بر محاسبات دقیق ریاضی برای تخمین و حذف پارامترهای مزاحم محیطی ارائه می دهد[۳]. این ایده را اولین بار *Moreno* در ۱۹۹۶ برای نگاشت بردارهای بازنمایی نویزی به تمیز به کار گرفت[۴] و از آن پس همواره یکی از مهم ترین زمینه‌های مورد علاقه در مباحث مربوط به بهبود کیفیت سیستم‌های بازشناسی مقاوم گفتار بوده است. تکنیک سری های تیلور برداری به دو طریق قابل استفاده است:

- اصلاح بردارهای بازنمایی و اعمال به مدل بازشناسی (مقاوم سازی بردارهای بازنمایی)
- اصلاح پارامترهای مدل بازشناسی (مقاوم سازی مدل)

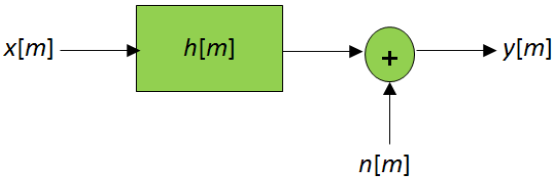
در روش اول ابتدا بردارهای بازنمایی جبران سازی شده و سپس بازشناسی می‌شوند، در حالی که در روش دوم بردارهای

اصلاح نشده و آلوده به نویز به مدل مقاوم شده اعمال می شوند. الگوریتم مورد استفاده برای هر دو روش کاملا تحلیلی بوده و نتایج حاصل از پیاده سازی آن ها برای جبران سازی نویز جمع شونده و کانال انتقال خطی بسیار قابل توجه می باشند[۵].

در مقاله حاضر، اصول و جزئیات الگوریتم سری های تیلور برداری برای اصلاح پارامترهای بازنمایی صوتی مورد بحث قرار می گیرد. در ابتدا نحوه مدل کردن تنوعات محیطی موجود بر روی سیگنال گفتار و معادلات مرتبط برای بیان ریاضی اثر این تنوعات نشان داده خواهد شد. سپس الگوریتم تشریحی و گام به گام جبران سازی اثر کانال خطی انتقال و نویز جمع شونده از روی بردارهای بازنمایی استخراج شده به روش $LFBE$ بیان گردیده و نحوه تخمین پارامترهای نویز و کانال، بدون وجود هیچ گونه اطلاعات قبلی و تنها از روی دادگان نویزی مشاهده شده به طور کامل شرح داده خواهد شد. در ادامه نیز نتایج آزمایشات عملی پیاده سازی الگوریتم بر روی یک سیستم بازشناسی گفتار نمونه استاندارد، برای مجموعه دادگان فارس دات میکروفنی (به ازای SNR های مختلف) و تلفنی (به ازای دو حالت انطباق و عدم انطباق) ارائه شده و با نتایج حاصل از حالت جبران نشده مقایسه خواهند گردید.

۲- توصیف مدل و مفروضات

مدل کلاسیک به کار رفته برای توصیف نحوه اثر گذاری عوامل مزاحم محیطی بر روی سیگنال گفتار مطابق شکل (۱) است.



شکل (۱): نحوه تاثیر تنوعات محیطی بر روی سیگنال گفتار

$x[m]$ و $y[m]$ به ترتیب نماینده سیگنال گفتار تمیز و تخریب شده در حوزه زمان و $n[m]$ و $h[m]$ نیز بیانگر اثر نویز جمع شونده و کانال انتقال هستند[۶].

نویز جمعی ناشناخته از طریق شیفت دادن میانگین طیفی و افزایش واریانس کل توزیع روی سیگنال تأثیر می‌گذارد. تغییر در کانال تلفن، تغییر در میکروفون یا اضافه شدن یک گوینده باعث اغتشاش‌های کانوالوی در سیگنال گفتار می‌شود. این اغتشاش‌ها یک تغییر جمعی در حوزه لگاریتم سیگنال ایجاد کرده و منجر به یک آفست متغیر با زمان می‌شوند[۷].

معادله (۱) رابطه بین طیف توان گفتار نویزی و گفتار تمیز را نمایش می دهد.

$$|Y(w)| = |X(w)|.|H(w)|^2 + |N(w)| \quad (۱)$$

که در آن $|X(w)|$، $|N(w)|$ و $|Y(w)|$ به ترتیب چگالی طیف توان گفتار تمیز، نویز و گفتار تخریب شده اند و $|H(w)|^2$ طیف توان کانال خطی انتقال گفتار است. معادله فوق در حوزه لگاریتم انرژی به صورت زیر بیان می شود[۴]:

$$y = x + h + log(1 + exp(n - x - h)) \quad (۲)$$

که در آن x و y بردار بازنمایی گفتار تمیز و نویزی و تابع $f(n,x,h) = h + log(1 + exp(n - x - h))$ تنوعات محیطی هستند. با یک اصلاح جزئی و افزودن ماتریس DCT می توان رابطه فوق را در حوزه کپستروم بازنویسی کرد. اما از آن جا که اصلاح ویژگی های $LFBE$ ساده تر است، در مقاله حاضر تکنیک VTS بر روی پارامترهای $LFBE$ انجام شده و سپس بردارهای بازنمایی حاصل به $MFCC$ تبدیل و به مدل بازشناسی اعمال می شوند.

برای تخمین مولفه های احتمالاتی گفتار تخریب شده، باید تخمینی از مدل توزیع احتمال گفتار تمیز تعلیمی در دست باشد. در این حالت می توان نگاشتی بین مولفه های نظیر گفتار تمیز و نویزی برقرار کرده و عوامل مزاحم محیطی را تخمین زد. تابع توزیع احتمال بردارهای بازنمایی گفتار تمیز تعلیمی به صورت مجموعی از توزیع های گوسی تکی یا به اصطلاح *Gaussian Mixture Model* فرض می شود[۸]:

$$p(x) = \sum_{k=1}^{K} p(k).\mathcal{N}(x;\mu_{x,k},\Sigma_{x,k}) \quad (۳)$$

در این رابطه K ، $p(k)$ ، $\mu_{x,k}$ و $\Sigma_{x,k}$ به ترتیب برابرند با تعداد کل توزیع های گوسی و احتمال اولیه، بردار میانگین و ماتریس کواریانس هر یک از توزیع ها. همچنین، تابع احتمال نویز جمعی با یک توزیع گوسی تکی مدل شده و کانال انتقال ناشناخته، اما خطی و تغییر ناپذیر با زمان فرض می گردد[۶]. اکنون فرض می شود که تاثیر مولفه های مزاحم بر روی گفتار تمیز، توزیع GMM آن را برهم نمی زند و می توان توزیع احتمال گفتار تخریب شده را نیز به صورت مجموعی از مولفه-های گوسی تکی فرض کرد[۹]:

$$p(z) = \sum_{k=1}^{K} p(k).\mathcal{N}(z;\mu_{z,k},\Sigma_{z,k}) \quad (۴)$$

هدف نهایی الگوریتم VTS تخمین عوامل مزاحم محیطی در طی فرایند نگاشت پارامترهای توزیع احتمال گفتار تمیز به نویزی است. با در اختیار داشتن پارامترهای تابع تنوعات می‌توان توزیع احتمال بردار بازنمایی اصلاح شده برای فریم های دیده نشده سیگنال گفتار را به کمک روش می نیمم میانگین مربعات خطا *(MMSE)* تخمین زد[۱۰]. بخش بعدی به توصیف نحوه این فرایند اختصاص دارد.

۳- الگوریتم سری های تیلور برداری

۳-۱- توصیف الگوریتم

با فرض وجود داده های در دسترس زیر:

۱- مجموعه ای ازبردارهای بازنمایی آلوده به نویز و اثر انتقال کانال

۲- تابع *PDF* گفتار تمیز

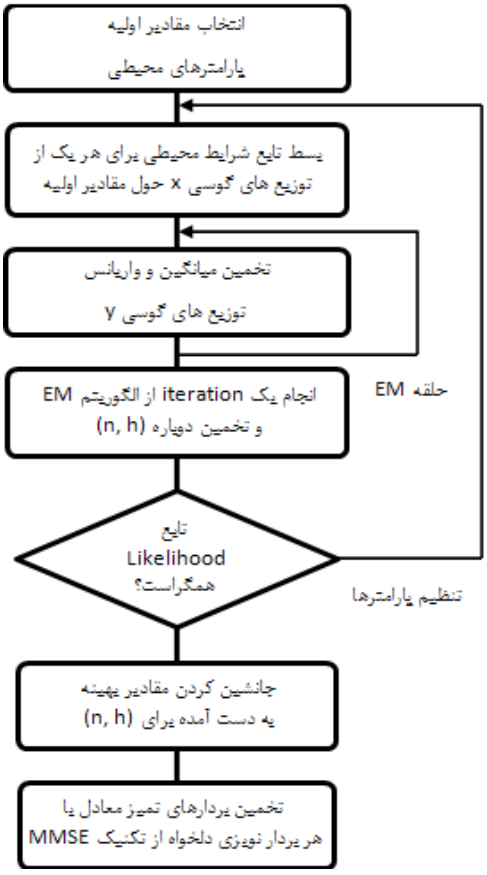
۳- مجموعه ای از مقادیر اولیه برای نویز و کانال شامل:

- میانگین نویز: می‌نیمم بردارهای نویزی دیده شده
- میانگین کانال: اختلاف بین میانگین بردارهای بازنمایی نویزی و تمیز
- ماتریس کواریانس بردارهای بازنمایی تمیز

و با در نظر گرفتن مدل فرض شده برای اثر تنوعات مزاحم محیطی، الگوریتم گام به گام جبران سازی پارامترهای محیطی از روی پارامترهای گفتار نویزی در حوزه لگاریتم طیف توان به ترتیب زیر تحقق می‌پذیرد[۱۱]:

- تعلیم مدل انتخابی *GMM* برای گفتار تمیز و تخمین پارامترهای تابع توزیع احتمال آن
- انتخاب مقادیر اولیه دسته پارامترهای مجهول محیطی $\{\mu_{n0}, \Sigma_{n0}, h_0\}$
- بسط تابع تنوعات محیطی برای هر یک از توزیع های گوسی توزیع احتمال گفتار تمیز حول $\{\mu_{n0}, \Sigma_{n0}, h_0\}$
-تخمین پارامترهای توزیع گفتار نویزی $(\mu_{y,k}, \Sigma_{y,k})$
- تکرار یک بار الگوریتم *EM* برای باز تخمین دسته مجهولات $\{\mu_n, \Sigma_n, h\}$
- در صورتی که تابع *Likelihood* همگرا بود، مقادیر بهینه برای $\{\mu_n, \Sigma_n, h\}$ به دست آمده اند. در غیر این صورت $\{\mu_n, \Sigma_n, h\}$ با $\{\mu_{n0}, \Sigma_{n0}, h_0\}$ جای گذاری شده و الگوریتم برای تکرار به مرحله ۲ باز می گردد.
- تخمین بردارهای بازنمایی تمیز از روی بردارهای بازنمایی نویزی با استفاده از مقادیر تقریب زده شده نویز و کانال به کمک تکنیک *MMSE*

شکل (۲) بلوک دیاگرام کلی استخراج بردارهای بازنمایی اصلاح شده به ازای هر یک از بردارهای ویژگی تخریب شده در حوزه لگاریتم طیف را نمایش می دهد[۱۲].



شکل(۲): بلوک دیاگرام الگوریتم اصلاح بردارهای بازنمایی به روش *VTS*

۳-۲- تخمین پارامترهای مزاحم محیطی

همان گونه که اشاره شد، توزیع احتمال گفتار نویزی نیز به صورت *GMM* در نظر گرفته می شود. میانگین و کواریانس این توزیع از روابط (۵) و (۶) به دست می آید.

$$\mu_{y,k} = (I + \nabla_x f)'\mu_{x,k} + (\nabla_n f)'\mu_n + (\nabla_h f)'h + g(n_0, x_0, h_0) \quad (۵)$$

$$\Sigma_{y,k} = (I + \nabla_x f)'\Sigma_{x,k}(I + \nabla_x f) + (\nabla_n f)'\Sigma_n(\nabla_n f) \quad (۶)$$

در روابط فوق $\mu_{y,k}$ ، $\Sigma_{y,k}$ و $\mu_{x,k}$ ، $\Sigma_{x,k}$ به ترتیب نماینده بردار میانگین و ماتریس کواریانس k امین توزیع گوسی گفتار تخریب شده و گفتار تمیز و μ_n ، Σ_n و h پارامترهای مجهول نویز و اعوجاج خطی هستند. $(\nabla_x f)'$، $(\nabla_n f)'$ و $(\nabla_h f)'$ نیز ترانهاده گرادیان تابع تنوعات محیطی به ترتیب نسبت به بردار گفتار تمیز، نویز و کانال اند و I بردار واحد هم مرتبه با بردارهای گرادیان است[۱۳].

برای تخمین دسته پارامترهای مجهول نیاز است تا تقریبی از احتمال پسین *(a posteriori probability)* تخصیص یافته به هر یک از توزیع های گوسی در دست باشد. به این منظور تابع چگالی احتمال گفتار تخریب شده به صورت رابطه

(۷) نمایش داده می شود.

$$P(y\mid k)=\sum_{k=1}^{K}P(k)P(y\mid k,\lambda)=\sum_{k=1}^{K}P(k)N(y;\mu_{y,k},\Sigma_{y,k}) \qquad (۷)$$

که در آن $p(k)$, λ , $\mathcal{N}(y;\mu_{y,k},\Sigma_{y,k})$ به ترتیب عبارتند از احتمال اولیه k امین توزیع گوسی، دسته پارامترهای مجهول و تابع احتمال گفتار نویزی به ازای k امین توزیع. حال با توجه به تخمین به دست آمده برای میانگین و کواریانس توزیع، می‌توان احتمال پسین هر یک از توزیع های مدل *GMM* را طبق رابطه (۸) محاسبه نمود[۱۴].

$$p(y|k)=\frac{p(k)p(y|k,\lambda)}{\sum_{k'=1}^{K}p(k')p(y|k',\lambda)} \qquad (۸)$$

تخمین پارامترهای تنوعات محیطی طی فرایندی تکراری انجام می شود. این فرایند با استفاده از الگوریتم *EM* سعی در حداکثر نمودن شباهت بین نمونه های مشاهده شده دارد. مقادیر بهینه پارامترها در هر بار تکرار الگوریتم تخمین با شرط *Maximum Likelihood* تعیین می شود. اما از آن جا که تخمین مولفه های مجهول با اعمال مستقیم الگوریتم *ML* دشوار است، به عنوان جایگزین آن از یک الگوریتم ماکزیمم تخمین استفاده می شود که به صورت تکراری پارامترهای نویز و اعوجاج کانال، و همچنین مولفه های گفتار نویزی را به روزآوری می کند[۱۳]. مقادیر بهینه برای میانگین و کواریانس نویز و بردار کانال در هر تکرار الگوریتم *EM* از روابط (۹) تا (۱۱) به دست می آیند[۱۴].

$$\hat{\mu}_n=\frac{\sum_{t=1}^{T}\sum_{k=1}^{K}p(k|y_t,\lambda)\mu_n(y_t,k,\lambda)}{\sum_{t=1}^{T}\sum_{k=1}^{K}p(k|y_t,k,\lambda)} \qquad (۹)$$

(۱۰)

$$\hat{\Sigma}_n=\frac{\sum_{t=1}^{T}\sum_{k=1}^{K}p(k\mid y_t,\lambda)[\Sigma_n(y_t,k,\lambda)+\mu_n(y_t,k,\lambda)\mu_n'(y_t,k,\lambda)}{\sum_{t=1}^{T}\sum_{k=1}^{K}p(k\mid y_t,k,\lambda)} - \hat{\mu}_n\mu_n' \qquad (۱۱)$$

$$\hat{h}=[\sum_{t=1}^{T}\sum_{k=1}^{K}p(k\mid y_t,\lambda)(\nabla_h f)(1+\nabla_x f)^{-1}\Sigma_{x,k}^{-1}(1+\nabla_x f)^{-1}(\nabla_h f)']^{-1}$$

پس از تقریب زدن مقادیر بهینه پارامترهای محیطی در یک بار تکرار، مقادیر جدیدی برای ۳ مجهول به دست آمده اند. شرط همگرایی الگوریتم تخمین آن است که این ۳ به تدریج به سمت مقادیری مشخص همگرا شوند. بنابراین در صورتی که این مقادیر در چند تکرار متوالی الگوریتم *EM* تغییر نکنند، تخمین به همگرایی رسیده و متوقف می شود. در غیر این صورت باید پس از هر تکرار، مقادیر اولیه میانگین و واریانس نویز و بردار اعوجاج با مقادیر جدید جانشین شده مجددا به آغاز الگوریتم بازگشت و مراحل پیش گفته را تا جایی تکرار نمود که نتیجه دلخواه حاصل شود.

در پایان فرایند تخمین به روش *EM* تکراری، دسته پارامترهای تابع تنوعات محیطی حاصل شده اند که با استفاده از آن ها، میانگین و کواریانس گفتار نویزی از روابط (۵) و (۶) به دست می آیند. اکنون آمادگی تخمین بردارهای بازنمایی گفتار مرجع از روی دادگان تخریب شده وجود دارد.

۳-۳- جبران سازی بردارهای نویزی

راه حل مناسب برای استخراج بردارهای بازنمایی تمیز از معادل های نویزی آن ها، استفاده از روش حداقل سازی میانگین مربعات خطا *(MMSE)* است. با توجه به معلوم بودن *PDF* گفتار نویزی، در حالت کلی رابطه (۱۴) برای استخراج بردارهای بازنمایی اصلاح شده نتیجه می شود.

$$\hat{x}_{MMSE}=E(x|y)=\int_X x\,p(x\,|y)dx \qquad (۱۲)$$

رابطه نهایی تخمین بردار بازنمایی گفتار تمیز به ازای هر یک از بردارهای ویژگی تخریب شده به صورت زیر خواهد بود[۱۴]:

$$\hat{x}_{MMSE}=y-\sum_{k=1}^{K}p(k|y)\,(\mu_{x,k}-\mu_{y,k}) \qquad (۱۳)$$

۴- نتایج آزمایشات

جهت ارزیابی کارایی الگوریتم ارایه شده در این مقاله، تکنیک *VTS* برای اصلاح بردارهای بازنمایی میکروفنی و تلفنی مجموعه دادگان فارس دات به کار رفته است. نرخ نمونه برداری سیگنال گفتار برای گویش های میکروفنی و تلفنی به ترتیب برابر ۸ و ۱۶ کیلو هرتز می باشد. بهینه ترین تعداد مولفه های گوسی *GMM* مورد استفاده برای مدل کردن گفتار تمیز برابر با ۱۲۸ اختیار شده است.

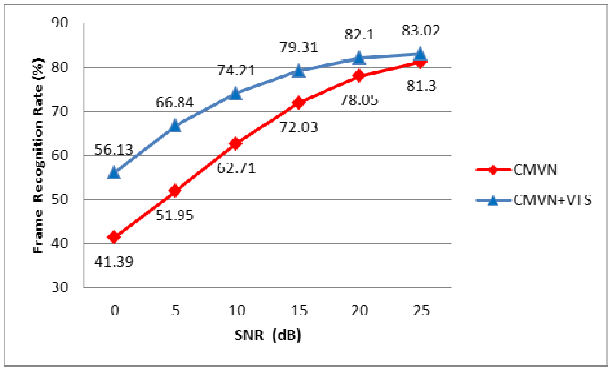
در حالت میکروفنی، مجموعه دادگان شامل ۴۰۰ جمله ادا شده توسط ۲۰۰ گوینده مختلف است که ۷۵٪ از آن ها برای تعلیم مدل و ۲۵٪ باقیمانده جهت آزمون اختصاص یافته اند. از ۳۰۰ جمله گفتار تمیز موجود، بردارهای بازنمایی ۳۹ بعدی (شامل ۱۲ ضریب *MFCC* به علاوه لگاریتم انرژی و مشتقات اول و دوم آن‌ها) استخراج شده و پس از نرمالیزه کردن بردارها به میانگین و واریانس، به مدل مرجع بازشناسی مبتنی بر شبکه عصبی*MLP* تعلیم داده می شوند.

۲۵٪ باقیمانده از دادگان گفتار تمیز به صورت دستی با نویز سفید گوسی در *SNR* های صفر ، ۵ ، ۱۰ ، ۱۵ ، ۲۰ و ۲۵ دسی‌بل مخلوط شده و بردارهای بازنمایی *LFBE* نویزی

استخراج می شوند. این بردارهای ویژگی به روش *VTS* اصلاح شده و به حوزه کپستروم انتقال می یابند. پس از استخراج پارامترهای *MFCC* شامل ۳۹ بعد و نرمالیزه کردن، هر دو دسته دادگان اصلاح نشده و اصلاح شده به روش *VTS*، برای تست به مدل بازشناسی اعمال می‌شوند.

مدل بازشناسی یک شبکه عصبی *MLP* تک لایه پنهان با ساختار نورونی < ۳۴-۱۰۰-۳۹×۷ > می‌باشد. در ورودی شبکه بردار بازنمایی فریم جاری به همراه بردارهای بازنمایی ۳ فریم مجاور چپ و راست آن (در مجموع ۷ فریم) قرار می‌گیرند. در لایه پنهان شبکه ۱۰۰ نورون و در خروجی شبکه نیز به تعداد آواهای موجود در دادگان تعلیمی یعنی ۳۴ نورون در نظر گرفته شده است. نورون های هر دو لایه دارای توابع غیرخطی از نوع تانژانت هایپربولیک هستند. مقادیر وزن های اولیه شبکه به صورت تصادفی و در محدوده ۱- تا ۱ انتخاب شده اند. 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 مومنتم نیز برابر ۰/۲ در نظر گرفته شده است.

شبکه مذکور چندین بار با مقادیر وزن های تصادفی متفاوت اولیه تعلیم داده شده و در نهایت نرخ صحت بازشناسی گفتار برای بردارهای بازنمایی تمیز برابر ۸۳/۹۳ درصد به ازای فریم های غیر سکوت به دست آمده است. در نمودار شکل (۳) نرخ بازشناسی بر اساس درصد صحت بازشناسی فریم‌های غیرسکوت برای دو دسته بردارهای بازنمایی نویزی و اصلاح شده و به ازای *SNR* های مختلف به صورت مقایسه ای بیان شده است.



شکل (۳): نتایج حاصل از بازشناسی بردارهای بازنمایی نویزی و جبران سازی شده میکروفنی به روش *VTS* به ازای *SNR* های مختلف

نتیجه مهم حاصل از این پیاده سازی، افزایش بیشتر نرخ بازشناخت در *SNR* های پایین است. متوسط افزایش نرخ بازشناسی برای *SNR* های بالا در حدود ۴/۲ درصد و برای *SNR* های پایین بیش از ۱۴ درصد است که مزیت بزرگی برای الگوریتم *VTS* محسوب می گردد.

برای دادگان گفتاری تلفنی واقعی، دو آزمون مجزا جهت ارزیابی بهبود ناشی از اعمال تکنیک *VTS* به کار می رود. در حالت انطباق دادگان تعلیم و تست، مجموعه آموزشی شامل همان ۳۰۰ جمله گفتار تمیز مورد استفاده در حالت میکروفنی است. با این تفاوت که پهنای باند فرکانسی داده های موجود به محدوده پاسخ فرکانسی کانال تلفن (۳۰۰ – ۳۴۰۰ *Hz*) تغییر یافته و فیلتر بانک مورد استفاده برای تولید بردارهای بازنمایی *LFBE* متشکل از ۱۵ فیلتر خواهد بود. با اعمال تبدیل کسینوسی گسسته و نرمالیزه کردن به میانگین و واریانس طیف، پارامترهای بازنمایی *MFCC* برای آموزش به شبکه آماده خواهند بود.

فارس دات شامل ۱۲۸ گویش تلفنی واقعی است که برای آزمون مدل به کار گرفته شده اند. پس از استخراج پارامترهای *LFBE* شامل ۱۵ بعد، اصلاح به روش *VTS* بر روی آن ها انجام شده و هر دو دسته دادگان اصلاح نشده و اصلاح شده به شبکه *MLP* پیش گفته اعمال می شوند.

اما در آزمون دوم، فرض می شود که تعلیم و تست هر دو باید با دادگان تلفنی واقعی در حالت انطباق صورت گیرند. به این منظور مجموعه دادگان فارس دات تلفنی اصلاح نشده و اصلاح شده به دو مجموعه تعلیم و آزمون تقسیم می شود، به طوری که ۴۸ گویش از هر جمله فارس دات به *Train* و ۱۶ گویش از هر جمله به *Test* اختصاص می یابد. در این حالت به ترتیب ۹۶ و ۳۲ جمله برای آموزش و ارزیابی سیستم وجود دارد که نسبت ۰/۲۵ برای نرخ آزمون را مشابه با حالات قبلی حفظ می کند. جدول (۱) نتایج مربوط به اصلاح و بازشناسی گفتار تلفنی را به صورت خلاصه بیان می کند.

جدول (۱): درصد بهبود بازشناسی ناشی از اصلاح به روش *VTS* در حالت انطباق و عدم انطباق دادگان آزمون و تست تلفنی

	عدم تطبیق	تطبیق
بدون اصلاح	۶۴/۱۴	۶۵/۷۰
اصلاح به روش *VTS*	۶۵/۲۵	۶۸/۲۶

افزایش نرخ بازشناسی حاصل از اعمال روش سری های تیلور برداری برای حالت عدم انطباق محیط تعلیم و تست بیش از ۱ درصد است. در حالی که میزان این افزایش در حالت انطباق دادگان آموزش و آزمون در حدود ۲/۵ درصد خواهد بود. تعبیر این نتیجه بدان معناست که استفاده از مدل بازشناسی تعلیم

داده شده بر روی دادگان تمیز فیلتر شده و تست آن با بردارهای بازنمایی تلفنی اصلاح شده به کمک تکنیک *VTS*، نتیجه به مراتب ضعیف تری نسبت به آموزش و تست مدل بازشناخت با گفتار تلفنی دارد. بنابراین در هنگام طراحی و پیاده سازی یک سیستم بازشناسی واقعی، بهتر است تعلیم سیستم با دادگان تلفنی واقعی انجام پذیرد.

۵- نتیجه گیری

با توجه به اهمیت مقاوم سازی سیستم‌های بازشناسی گفتار در برابر تنوعات محیطی موجود در کاربردهای عملی، و نظر به عدم کارایی روش‌های متداول استخراج ویژگی مانند *MFCC* در برابر این تنوعات، در این مقاله تکنیک تحلیلی و مبتنی بر فرمول بندی دقیق ریاضی سری های تیلور برداری *(VTS)* برای اصلاح و جبران سازی بردارهای بازنمایی صوتی ارائه گردید. نتایج حاصل از پیاده سازی این تکنیک بر روی مجموعه دادگان گفتاری تخریب شده با نویز سفید که اثر کانال انتقال را نیز در قالب تنوعات میکروفون و گوینده دارا می باشد، به خوبی نشان دهنده تاثیر آن در حذف اثر نویز جمع شونده و کانال انتقال خطی از روی سیگنال گفتار است. به گونه ای که بهبود درصد بازشناسی به ویژه در *SNR* های پایین بسیار قابل توجه است.

نتیجه اصلی پروژه حاضر به هنگام محک زدن الگوریتم معرفی شده بر روی گفتار تلفنی واقعی مشخص می شود. هر دو حالت استفاده از دادگان تمیز یا تلفنی اصلاح شده برای آموزش مدل صوتی، منجر به بهبود کیفیت بازشناسی می گردد. اما بهره بردن از خود بردارهای بازنمایی اصلاح شده تلفنی به جای بردارهای بازنمایی گفتار تمیز برای تعلیم مدل بازشناسی همواره ارجحیت دارد.

۶- مراجع

[۱] منصور ولی، "بازشناسی مقاوم گفتار به منظور جبران سازی تنوعات گفتار میکروفنی و تلفنی توسط شبکه‌های عصبی"، پایان نامه دکتری، دانشگاه صنعتی امیرکبیر، بهار ۱۳۸۵.

[2] Jasha Droppo, “Noise Robust Automatic Speech Recognition”, Microsoft Research Center, pp, 12-43, 2008

[3] Yong Zhao and Bing-Hwang Juang, “On Noise Stimation for Robust Speech Recognition using Vector Taylor Series”, ICASSP , pp, 1-4, 2010.

[4] Michael L. Seltzer and Alex Acero, “HMM adaptation using Linear Spline Interpolation with Integrated Spline Parameter Training for Robust Speech Recognition”, INTERSPEECH, pp, 1-4, 2010.

[5] Pedro J.Moreno, “Speech Recognition in Noisy Environments”, Thesis, pp, 79-104, 1996

[6] Pedro J. Moreno, Bhiksha Raj and Richard M. Stern, “A Vector Taylor Series Approach for Environment-Independent Speech Recognition”, Microsoft Research Center, pp, 1-4, 2008.

[7] R. C. van Dalen and M. J. F. Gales, “Extended VTS for Noise-Robust Speech Recognition”, InterSpeech, pp, 1-4, 2009.

[8] O. Kalinli, M. L. Seltzer and A. Acero, “Noise Adaptive Trainig Using Vector Taylor Series Approach for Noise Robust Automatic Speech Recognition”, Microsoft Research Center, pp, 1-3, 2009.

[9] Nam Soo Kim, Do Yeong Kim, Byung Goo Kong and Sang Ryong Kim, “Application of VTS to environment Compansation with Noise Statistics”, ICASP, pp, 1-4, 1999.

[10] R. C. van Dalen and M. J. F. Gales, “Covariance Modeling for Noise Robust Speech Recognition”, Interspeech, pp, 1-4, 2008.

[11] Chandra Kaut Raut, Takuya Nishimoto, Shigeki Sagayama, “Model Composition by Lagrange Polynomial Approximation for Robust Speech Recognition in noisy Environment”, Interspeech, pp, 1-4, 2008.

[12] Zhao Xianyu, Ou Zhijian and Wang Zuoying, “Using Vector Taylor Series with Noise Clustering for Speech Recognition in non-stationary Noisy Environments”, High Technology Letters, pp, 1-5, 2006.

[13] Do Yeong Kim, Chang Kwan Un and Nam Soo Kim, “Speech Recognition in Noisy Environments Using First-Order Vector Taylor Series”, Speech Communication, pp, 1-4, 1998.

[14] Do Yeong Kim, Chang Kwan Un and Nam Soo Kim, “Speech Recognition in Noisy Environments Using First-Order Vector Taylor Series”, Speech Communication, pp, 1-4, 1998.