

روشی جدید در بازشناسی مقاوم گفتار مبتنی بر دادگان مفقود با استفاده از شبکه عصبی دوسویه

حجت محمدنژاد^۱، منصور ولی^۲

^۱ دانشگاه شاهد، دانشکده فنی

^۲ دانشگاه شاهد، دانشکده فنی و دانشگاه صنعتی خواجه نصیرالدین طوسی، دانشکده برق و کامپیوتر

چکیده

عملکرد سامانه‌های بازشناسی گفتار زمانی که گفتار توسط نویفه تخریب شده باشد، به شدت کاهش می‌یابد. یکی از روش‌های رایج برای مقاوم‌سازی سامانه‌های بازشناسی گفتار استفاده از روش دادگان مفقود است. در این روش مؤلفه‌هایی از نمایش زمانی-فرکانسی گفتار (اسپکتروگرام) که نسبت سیگنال به نویفه (SNR) آنها از یک آستانه کمتر است، به عنوان مؤلفه‌های نامعتبر یا مفقود برجسته‌دهی می‌شوند. این مؤلفه‌ها با استفاده از مؤلفه‌های معابر و اطلاعات آماری نسبت به دادگان تمیز، تخمین زده شده و جایگزین می‌شوند.

در این مقاله الگوی ویژگی‌های مفقود با دیدگاهی نو، به عنوان مستقله جبران‌سازی دادگان مطرح می‌شود. بدین صورت که با استفاده از شبکه عصبی دوسویه و انجام یک سری پردازش‌های غیرخطی و دوطرفه (جلوسو و برگشتی) از داشت نهفته در مدل، ناشی از یادگیری همزمان گفتار تمیز و نویفه بهره گرفته، بردارهای بازنمایی گفتار در جهت افزایش صحبت بازشناسی آواها بهبود می‌یابند.

در این روش نیازی به شناسایی مؤلفه‌های مفقود که یک بحث چالش برانگیز در حوزه بازشناسی مقاوم گفتار مبتنی بر دادگان مفقود است، نمی‌باشد؛ بلکه بازسازی در جهت هرچه شبیه‌تر شدن نمایی مؤلفه‌ها، خواه معابر باشد، خواه نامعتبر به مؤلفه‌های گفتار تمیز صورت می‌گیرد و این یک برتری بسیار چشم‌گیری است که در این تحقیق حاصل شده است. نتایج مقایسه این دو روش نشان می‌دهد که با استفاده از روش دادگان مفقود، $4/2$ درصد بهبود برای صحبت بازشناسی گفتار نویفه‌ای با نسبت سیگنال به نویفه $4dB$ حاصل شده در حالی که با استفاده از روش مبتنی بر شبکه عصبی دوسویه، $8/5$ درصد بهبود برای همان نسبت سیگنال به نویفه به دست آمده است.

واژگان کلیدی: بازشناسی مقاوم گفتار، دادگان مفقود، شبکه عصبی دوسویه

طرح شده است که در آن لازم است سامانه بازشناسی گفتار، به گونه‌ای منعطف و تطبیق‌پذیر طراحی گردد که امکان شناسایی گفتار نویفه‌ای و حصول به نرخ‌های بالای بازشناسی در شرایط محیطی متفاوت امکان‌پذیر باشد.
(Acero, 1993)

در بازشناسی مقاوم گفتار برای بهبود تطابق بین دادگان آموزشی و دادگان آزمون به طور معمول دو رویکرد کلی وجود دارد. در رویکرد اول، جبران اثر نویفه در دادگان آزمون، با اصلاح بُردارهای بازنمایی گفتار صورت می‌پذیرد که آنها را روش‌های مبتنی بر جبران‌سازی دادگان می‌نامند. در رویکرد دوم هدف، اصلاح پارامترهای مدل طبقه‌بندی کننده برای جبران اثر نویفه است که روش‌های جبران‌ساز طبقه‌بندی کننده، نامیده می‌شوند.

۱- مقدمه

با رشد روزافزون استفاده از سامانه‌های بازشناسی گفتار در کاربردهای عملی و روزمره، نیاز به حفظ کارایی بازشناسی گفتار در محیط‌های واقعی به عنوان امری اجتناب‌ناپذیر مطرح است. بنابراین هنگامی که از سامانه بازشناسی گفتاری که در محیط آزمایش‌گاهی آموخته شده است، در محیط واقعی استفاده شود اغلب کارایی سیستم بازشناسی، بهدلیل عدم انطباق^۱ دادگان آموزشی و دادگان محیط واقعی، به مقدار زیادی کاهش می‌یابد. از این‌رو بحث مقاوم‌سازی^۲ در برابر نویفه به عنوان یکی از ضرورت‌های زمینه‌های فعال تحقیقاتی در زمینه بازشناسی مقاوم گفتار در سال‌های اخیر

^۱- Mismatch

^۲- Robustness

روش، نواحی با SNR پایین، از اسپکتروگرام گفتار پاک شده و با مقادیر تخمین بهینه‌ای از مقادیر مورد قبولشان جایگزین می‌شوند.

مزیت روش ویژگی‌های مفقود در این است که در این روش، هیچ فرضی درباره ایستان بودن نوفة مخرب در نظر گرفته نمی‌شود؛ همچنین این روش، نیازی به اطلاعات کامل از ساختار طیفی نوفة نداشته و تنها به توصیفی از نواحی نمایش زمانی- فرکانسی، با عنوان معتبر و نامعتبر بسنده *Bourlard and Dupont, 1996; Cooke et al., 1997; Moore, 1994*) روش مبتنی بر ویژگی‌های مفقود، از دو بخش اصلی شناسایی مؤلفه‌های مفقود و اصلاح آنها تشکیل شده است.

ایده کارآمد دیگری که در حوزه اصلاح بُردارهای بازنمایی گفتار مطرح شده، استفاده از شبکه عصبی دوسویه^۹ (BNN) است که به صورت هم‌زمان بر روی دادگان تمیز و نوفة‌ای تلفنی، آموزش داده شده است. در این روش پس از انجام یک سری پردازش‌های غیرخطی و دوطرفه (جلوسو و برگشتی) از دانش نهفته در مدل، ناشی از یاد گرفتن گفتار تلفنی و نوفة‌ای بهره گرفته و بُردارهای بازنمایی گفتار، در جهت افزایش صحّت بازشناسی آواهای گفتار بهبود داده شده‌اند (*Lippmann and Carlson, 1997*).

نقطه قوت روش اصلاح ویژگی مبتنی بر شبکه عصبی دوسویه نسبت به روش ویژگی‌های مفقود، این است که نیازی به شناسایی مؤلفه‌های مفقود ندارد و بازسازی را در جهت هرچه شبیه‌تر شدن تمامی مؤلفه‌های معتبر یا نامعتبر، به مؤلفه‌های گفتار تمیز صورت می‌دهد و این بک برتری بسیار چشم‌گیری است که در این تحقیق حاصل شده است؛ چرا که در عمل، بحث شناسایی مؤلفه‌های مفقود، که یک بحث چالش‌برانگیز را در تمامی روش‌های بازشناسی مقاوم گفتار مبتنی بر ویژگی‌های مفقود است و ارتباط مستقیمی با میزان صحّت بازشناسی دارد، حذف می‌کند.

در این تحقیق، روش ویژگی‌های مفقود برای اصلاح بُردارهای بازنمایی گفتار، بر روی دادگان گفتار فارسی آگشته به نوفة در سطوح مختلف پیاده‌سازی می‌شود و هم‌چنین الگوی ویژگی‌های مفقود با استفاده از شبکه عصبی دوسویه در ادامه تحقیقات قبلی ما به عنوان مسئله جبران‌سازی دادگان، مطرح خواهد شد که در آن عناصر مفقود در حوزه طیف گفتار، به گونه‌ای بازسازی می‌شوند که حاوی اطلاعات مفید در جهت بازشناسی آواهای گفتار باشند؛ بدون آنکه

^۹-Bidirectional Neural Network (BNN)

در روش‌های جبران‌سازی دادگان از قبیل *Normalizer کردن کپستروم وابسته به کد^۱ (CDCN)، (Moreno, 1996)، سری‌های تیلوربرداری^۲ (VTS)، Porter and Boll, (Boll, 1979) و فیلتر وینر (Porter and Boll, 1984)، اثر نوفة روی دادگان، با تخمین طیف نویز جبران می‌گردد و سایر روش‌ها از قبیل جبران‌سازی کپستروم مبتنی برتابع گوسی چندمتغیره^۳ (Moreno, 1996) و فیلتر Neumeyer and Weintraub, (POF) (Neumeyer and Weintraub, 1994) به تطبیق بین دادگانی که به طور همزمان در شرایط آموزش و آزمون ضبط شده‌اند، می‌پردازند.*

روش‌های جبران‌سازی طبقه‌بندی‌کننده از قبیل ترکیب مدل موازی^۴ (PMC) (*Gales and Young, 1993*) و ترکیب مدل^۵ (*Varga and Moore, 1990*)، توزیع کلاس‌های آولی را برای رسیدن به نوفة جمعی، اصلاح می‌کنند و سایر روش‌ها از قبیل رگرسیون خطی با بیشترین احتمال^۶ (*Leggetter and Woodland, 1994*) (MLLR)، مشخصات توزیع‌ها را برای متناسب‌شدن با گفتار نوفة‌ای آزمون، تغییر می‌دهند. یکی از موانع تمامی این روش‌ها در این است که همگی آنها، نوфе مورد نظر را ایستان فرض می‌کنند. تمامی این روش‌ها در مواجهه با نوфе ایستان با توان پایابن یا متوسط، موفقیت به خوبی از خود نشان می‌دهند (گفتار نوفة‌ای با $SNR = 10 \text{ dB}$ یا بالاتر)، ولی این روش‌ها در مواجهه با نویز با سطوح بالاتر، کارآمدی خود را از دست داده و حتی در برخورد با نویز غیرایستان به طور کامل بی‌نتیجه خواهند بود (*Raj et al., 1997*).

یکی از روش‌های مؤثر و کارآمد که در حوزه جبران‌سازی دادگان آگشته به نوفة ایستان و غیرایستان مطرح شده است، روش ویژگی‌های مفقود است (*Hennansky et al., 1996; Moore, 1997*). در این روش برای تخمین مؤلفه‌های مفقود اسپکتروگرام، از همبستگی بین مؤلفه‌های طیفی گفتار در طول گفتار، استفاده می‌شود. در روش‌های مبتنی بر ویژگی‌های مفقود، نه تنها می‌توان مقدار SNR را در باند‌های فرکانسی متمایز، متفاوت دانست، بلکه می‌توان تخریب سیگنال گفتار در طول زمان را نیز تحت تأثیر سطوح مختلف نوфе در نظر گرفت. در این

¹codeword dependent cepstral normalization

²vector Taylor series

³ spectral subtraction

⁴ multivariate Gaussian based cepstral compensation

⁵ probabilistic optimal filtering

⁶ parallel model combination

⁷ model composition

⁸ maximum likelihood linear regression



کاربرد دارد (Moreno, 1996). در این تحقیق به شرح این تکنیک تنها در تخمین نوافه بردارهای بازنمایی تخریب شده توسط نوافه جمعی از روی سیگنال گفتار فارسی خواهیم پرداخت. بدین ترتیب که ابتدا بردارهای بازنمایی لگاریتم انرژی بانک فیلتر^۱ (LFBE) از گفتار نوافه ای استخراج شده و سپس به کمک تکنیک VTS میانگین نوافه جمعی برای تک تک باندهای فرکانسی به دست می آید.

سری تیلور برداری، یک الگوریتم جبران ساز نوافه است که به دنبال تخمین حداکثر شباهت از پارامترهای فیلتر خطی و نوافه جمعی بر روی بردارهای لگاریتم طیفی سیگنال های نوافه ای است (Moreno, 1996). اگر $Y(m)$ بیانگر بردار لگاریتم طیفی فریم m یک گویش باشد که توسط فیلترینگ خطی و نوافه جمعی تخریب شده باشد و همچنین $X(m)$ بیان گر سیگنال تمیز و عاری از نوافه همان گویش باشد؛ در این صورت می توان رابطه (۱) را بین این دو تعریف کرد (Moreno, 1996):

$$Y(m) = X(m) + H + \log(1 + \exp(N - H - X(m))) \quad (1)$$

که در آن H ، لگاریتم مربع دامنه طیف پاسخ ضربه فیلتر خطی و N ، لگاریتم طیف نوافه محسوب می شود. فرض بر این است که نوافه، ایستان بوده و تفاوت در بردارهای طیفی مجزای تخریب شده نوافه ای، مربوط به تفاوت های تحقق یک فرایند گوسی باشد. همچنین فرض بر این است که توزیع لگاریتم طیف نوافه در بردارهای مختلف، گوسی بوده و می توان نوافه را تنها با دو مؤلفه میانگین Σ_N و واریانس Σ_N نمایش داد.

توزیع لگاریتم طیف سیگنال تمیز هم به عنوان یک فرایند گوسی مختلط فرض می شود. مجموعه پارامترهای مختلط گوسی Φ نیز از مجموعه دادگان گفتار تمیز به دست می آید

مسئله موجود در VTS در چارچوب رابطه (۱)، برآورد احتمال حداکثر از پارامتر کانال H ، میانگین Σ_N و واریانس Σ_N از طیف نوافه می باشد. در حقیقت اگر مجموعه ای از بردارهای طیفی سیگنال های نوافه ای را با Y نمایش دهیم، در روش VTS ما به دنبال حداکثر شباهت بین پارامترهای نوافه سیگنال های نوافه ای و پارامتر کانال H ، میانگین Σ_N و واریانس Σ_N از طیف نوافه هستیم. این برآورد را می توان به صورت رابطه (۲) به نمایش گذاشت:

^۱ - Logarithm of Filter-Bank Energies (LFBE)

نیاز به شناسایی دقیق مؤلفه های مفقود طیف باشد. درنهایت دو روش با هم ترکیب شده و به صورت متوالی در اصلاح بردارهای بازنمایی گفتار نوافه ای ارزیابی خواهد شد. نتیجه طراحی چنین تکنیک هایی مقاوم تر شدن سامانه بازناسی گفتار و حصول به نرخ های بالاتر بازناسی در شرایط واقعی خواهد بود.

در ادامه در (بخش های دو و سه) مقاله، به ترتیب روش اصلاح بردارهای بازنمایی گفتار مبتنی بر ویژگی های مفقود و شبکه عصی دوسویه مطرح خواهد شد. در بخش چهار دادگان مورد استفاده در این تحقیق و نحوه استخراج ویژگی از آنها خواهد آمد. سپس در بخش پنج مقاله، آزمایش ها و نتایج حاصل از آنها را به خواهد شد و درنهایت در بخش شش، تحلیل نتایج صورت خواهد پذیرفت.

۲- روش ویژگی های مفقود در جبران سازی بردارهای بازنمایی نوافه ای

روش ویژگی های مفقود از دو بخش اصلی تشکیل می شود. بخش اول شناسایی مؤلفه های مفقود طیف گفتار و بخش دوم بازسازی این مؤلفه های مفقود از روی بخش های سالم گفتار است. در ادامه به تشریح کامل هر یک از این دو بخش خواهیم پرداخت.

۲-۱- بخش اول: شناسایی مؤلفه های مفقود

یکی از مشکل ترین و حساس ترین بخش های سامانه های بازناسی مقاوم گفتار مبتنی بر ویژگی های مفقود، تخمین ماسک های اسپکترو گرافیک است که مؤلفه های نامعتبر طیف را شناسایی می کند (Raj and Stern, 2005). این تخمین می تواند به چندین روش انجام شود، ممکن است SNR هر مؤلفه طیف، به صورت مجزا برای شناسایی مؤلفه های مفقود تخمین زده شود و یا اینکه مؤلفه های مفقود، به طور مستقیم با استفاده از معیار دیگری غیر از SNR، تفکیک شوند. در این تحقیق با تخمین هر مؤلفه طیف با استفاده از روش سری تیلور برداری، مؤلفه های نامعتبر شناسایی خواهد شد.

تکنیک سری تیلور برداری^۱ (VTS) از جمله کارترین روش های بازناسی مقاوم گفتار است که در هر دو حوزه اصلاح بردارهای بازنمایی و اصلاح مدل صوتی بازناسی

^۱ - Vector Taylor Series (VTS)

کواریانس بین مؤلفه‌های بردارهای بازنمایی طیفی) نمایش داده می‌شوند.

μ_k نشان‌دهنده میانگین k امین باند فرکانسی از m امین فریم بردار طیفی $(X(m, k), X, k)$ است. کواریانس بین $X(m, k_1)$ و $X(m, k_2)$ امین باند فرکانسی از m + ξ + k_2 امین بُردار طیفی $(X(m + \xi, k_2), X, k_2)$ نیز با $c(\xi, k_1, k_2)$ نمایش داده می‌شود که مقدار نرمالیزه شده نظری آن $r(\xi, k_1, k_2)$ است که از طریق روابط زیر به دست می‌آید:

(۴)

$$\alpha(k) = E[X(m, k)]$$

$$c(\xi, k_1, k_2) = E[(X(m, k_1) - \alpha_{k_1})^T (X(m + \xi, k_2) - \alpha_{k_2})]$$

$$r(\xi, k_1, k_2) = \frac{c(\xi, k_1, k_2)}{\sqrt{c(\xi, k_1, k_1)c(\xi, k_2, k_2)}}$$

در رابطه فوق [۴] عمل گر امید ریاضی را نشان می‌دهد. این پارامترها از بردارهای طیفی استخراج شده از مجموعه دادگان تمیز به دست می‌آید.

برای تخمین مؤلفه‌های طیفی نامعتبر در m امین فریم بردار طیفی $(X_u(m), X_r(m))$ گرد می‌آوریم؛ سپس همه مؤلفه‌های طیفی معتبر در نمایش اسپکتروگرام را که دارای کواریانس نرمالیزه شده حداقل ۰/۵ با حداقل یکی از مؤلفه‌های بردار $X_u(m)$ باشند، شناسایی و در بردار $(X_r(m))$ مرتب می‌کنیم. مقدار میانگین و کواریانس بین مؤلفه‌های $(X_u(m))$ و $(X_r(m))$ همچنین کواریانس متقابل بین آنها به وسیله پارامترهای میانگین و کواریانس حاصل از داشت اولیه فرایند گوسی مورد نظر ساخته می‌شوند.

نظر به اینکه فرایند گوسی فرض شده است، مقدار اصلاح شده مؤلفه‌های نامعتبر $(X_u(m))$ را می‌توان به وسیله الگوریتم MAP تخمین زد که در زیر رابطه نهایی آن آورده شده است.

$$\hat{X}_u(m) = \alpha_u + C_{ru} C_{rr}^{-1} (Y_r(m) - \alpha_r) \quad (5)$$

که در آن α_u و α_r به ترتیب نشان‌دهنده مقدار میانگین مؤلفه‌های $(X_u(m))$ و $(X_r(m))$ و همچنین C_{ru} به ترتیب کواریانس و کواریانس متقابل بین مؤلفه‌های $(X_u(m))$ و $(X_r(m))$ هستند. $\hat{X}_u(m)$ نیز معرف مقادیر بازسازی شده بردار ویژگی‌های مفقود نظری $(X_u(m))$ می‌باشد.

برای بازناسی بهتر لازم است مقادیر بازسازی شده ویژگی‌های مفقود در یک حیطه خاص، محدود شوند؛ که آن را در ترکیب با رابطه فوق در اصطلاح *Bounded MAP*

$$H, \alpha_N, \Sigma_N = \arg \max_{H, \alpha, \Sigma} \left\{ P(Y | H, \alpha, \Sigma, \Phi) \right\} \quad (2)$$

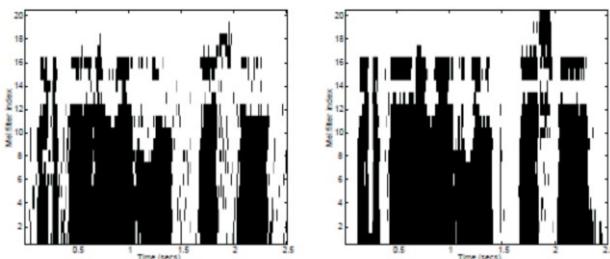
هنگامی که H, α_N و Σ_N تخمین زده شد، $(Y(m))$ را می‌توان از (H, α_N, Σ_N) با استفاده از تکنیک کمینه‌سازی میانگین مربعات خطای MMSE تخمین زد.

مقدار میانگین لگاریتم طیفی α_N را می‌توان به عنوان مقدار واقعی لگاریتم طیفی نویه نیز در نظر گرفت و از آن برای تخمین SNR محلی منحنی اسپکتروگرام سیگنال تخریب شده استفاده کرد.

اگر $(Y(m, k))$ بیان گر مقدار مؤلفه k ام باند فرکانسی از بردار طیفی m ام اسپکتروگرام نویه باشد و همچنین اگر k امین مؤلفه فرکانسی با α_N نمایش داده شود، مقدار تخمینی SNR از رابطه (۳) بدست خواهد آمد:

$$SNR(m, k) = \frac{Y(m, k) - \alpha_N(k)}{\alpha_N(k)} \quad (3)$$

همان‌طور که در بخش قبل شرح داده شد، ماسک اسپکتروگرافیک بر مبنای SNR تخمینی محاسبه می‌شود؛ بدین صورت که تمامی مؤلفه‌های اسپکتروگرام که مقدار $SNR(m, k)$ آنها، از مقدار آستانه T کمتر باشد، به عنوان



(شکل ۱): (سمت چپ) ماسک اسپکتروگرافیک به دست آمده از VTS برای یک گفتار نویه تخریب شده با $SNR = 10 dB$ (سمت راست) ماسک اسپکتروگرافیک از پیش دانسته برای همان گویش مفقود برچسبدهی می‌شوند.

(شکل ۱) ماسک اسپکتروگرافیک تخمین زده شده با استفاده از روش VTS در کنار ماسک اسپکتروگرافیک از پیش دانسته را که از روی اطلاعات دقیق سیگنال نویه افروده شده به گفتار به دست آمده است را نشان می‌دهد.

۲-۲- روشن بازسازی مبتنی بر کواریانس

یکی از روش‌های مبتنی بر بازسازی ویژگی‌های مفقود گفتار، روش بازسازی مبتنی بر کواریانس است (Raj, 2000). در این روش بردارهای لگاریتم طیفی استخراج شده از گفتار، نمونه‌های یک فرایند تصادفی گوسی ایستان فرض می‌شوند. داشت اولیه درباره سیگنال‌های گفتار تمیز، به صورت پارامترهای آماری (مقدار میانگین بردارها و



خروجی شاخه بازگشتی، خود یک بردار N_I مؤلفه‌ای است که با مؤلفه‌های نظیر خود از بردارهای بازنمایی ورودی u به صورت رابطه (۷) ترکیب می‌شوند.

$$x_i[n] = \lambda u_i + \sum_{l=0}^{N_r} r_l[n] w_{li}^r \quad i = 1, 2, \dots, N_I \quad (7)$$

رابطه (۷) نشان می‌دهد که حاصل جمع کسری $\lambda < 1$ از بردار بازنمایی u ، و یک ترکیب خطی از مقادیر لایه مخفی بازگشتی شبکه r ، ورودی شبکه بهازی این بردار بازنمایی را در دوره n آم تعلیم تشکیل می‌دهد. در ابتدای تعلیم $n = 1$ ، ورودی‌های شبکه همان بردارهای بازنمایی اصلی گفتار هستند ($x_i[1] = u_i$). در اولین دور تعلیم، مقادیر لایه مخفی بخش جلوسوی، بهازی هر ورودی حاصل می‌شوند که از آنها برای ورودی بخش بازگشتی در دور بعد تعلیم استفاده می‌شوند. لازم به ذکر است که نرمالیزه‌بودن بردارهای بازنمایی ورودی شبکه عصبی دوسویه یعنی مقادیر r_i ها بین $[-1, 1]$ در کنار خروجی‌های لایه مخفی بخش بازگشتی شبکه که در محدوده $[-1, 1]$ هستند، برای هم‌گراشدن رابطه (۷) ضروری است.

مقادیر لایه مخفی بازگشتی شبکه، مطابق رابطه (۸) خود به صورت یکتابع غیرخطی از روی مقادیر لایه مخفی بخش جلوسوی شبکه (y)، حاصل از دوره $1 - n$ تعلیم به دست می‌آیند. بدین ترتیب لازم است که برای کلیه دادگان تعلیم مقادیر y نظیر آنها از دوره فعلی به منظور استفاده در دوره بعدی تعلیم شبکه نگهداری شوند.

$$r_i[n] = f\left(\sum_{j=0}^{N_H} y_j[n-1] v_{jl}^r\right) \quad l = 1, 2, \dots, N_r \quad (8)$$

$$y_j[n] = f\left(\sum_{i=0}^{N_I} x_i[n] w_{ij}\right) \quad j = 1, 2, \dots, N_H \quad (9)$$

$$z_k[n] = f\left(\sum_{j=0}^{N_H} y_j[n] v_{jk}\right) \quad k = 1, 2, \dots, N_O \quad (10)$$

هدف از به کارگیری مسیر بازگشتی در شبکه BNN 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 مربوط به این است که شبکه جلوسوی برای هر آوای گفتار، بردارهای بازنمایی سالم از گفتار تمیز را در اختیار دارد که به علت همبستگی بهتر و تعداد بیشتر آنها نسبت به بردارهای بازنمایی تخریب شده و پراکنده نویزی، آنها را بهتر یاد می‌گیرد. علاوه بر آن، دانش یادگرفته شده در لایه مخفی شبکه، بیشتر حاوی اطلاعات آوازی گفتار است و بسیاری از پراکنده‌گی‌های بین ویژگی‌های ورودی که برای بازناسی آواها مفید نمی‌باشند نیز در آن

می‌نامند (Raj, 2000). رابطه (۶) نحوه اعمال این محدودیت را نشان می‌دهد.

$$\hat{X}_u(m, k) = \begin{cases} \hat{X}_u(m, k) & \text{if } \hat{X}_u(m, k) \leq X_u(m, k) \\ X_u(m, k) & \text{otherwise} \end{cases} \quad (6)$$

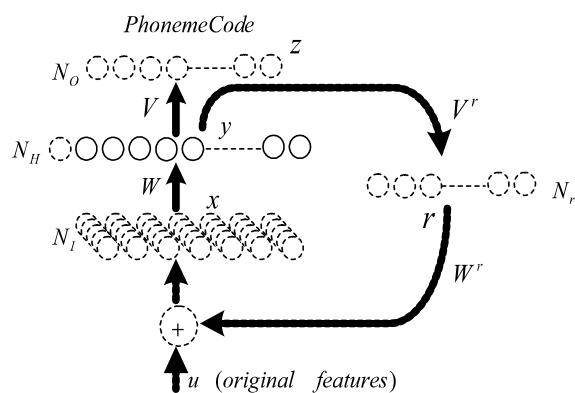
از آنجایی که سیگنال نوشهای در اثر اضافه شدن نوشه مخرب به گفتار تمیز، حاصل شده است، درنتیجه در رابطه (۶) مقادیر لگاریتم طیفی سیگنال گفتار نوشهای به عنوان یک کران بالا برای مقادیر بازسازی شده در نظر گرفته شده است.

۳- شبکه عصبی دوسویه و اصلاح بردارهای بازنمایی گفتار

شبکه عصبی دوسویه برای اوتین بار در سال ۲۰۰۶ پیشنهاد شده و برای اصلاح بردارهای بازنمایی لگاریتم طیفی گفتار تلفنی مورد استفاده واقع شد (Vali et al., 2006a). در این تحقیق از این روش برای اصلاح بردارهای بازنمایی گفتار نوشهای با نسبت‌های مختلف SNR استفاده می‌شود و با روش‌های رایج دادگان مفقود مقایسه خواهد شد. لذا در ابتدا ساختار و نحوه عملکرد این نوع شبکه ارائه می‌شود.

۳-۱- ساختار شبکه عصبی دوسویه

ساختار شبکه عصبی دوسویه مطابق (شکل ۲) مشتمل بر یک شبکه جلوسوی MLP با توابع غیرخطی تانژانت هایپربولیک در لایه مخفی، به همراه یک شاخه بازگشتی خود 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 است.



(شکل ۲): ساختار شبکه عصبی دو سویه

۲-۳- اصلاح بردارهای بازنمایی توسط شبکه دوسویه

همان طور که اشاره شد شبکه عصبی دوسویه (شکل ۲) از دو بخش اصلی تشکیل یافته است: یکی بخش جلوسو که شبیه به یک شبکه عصبی MLP معمولی است و کار دسته‌بندی الگوها را انجام می‌دهد و دیگری بخش بازگشتی شبکه است که از لایه مخفی بخش جلوسوی شبکه به ورودی شبکه بر می‌گردد و یک نگاشت از روی مقادیر لایه مخفی بخش جلوسو به فضای بردارهای بازنمایی ورودی فراهم می‌کند. این نگاشت برای تخمین یک بردار اصلاحی است که با کسری از بردار بازنمایی ورودی جمع شده و ورودی اصلاح شده به شبکه را فراهم می‌کند.

پس از اطمینان از هم‌گرایی شبکه، اکنون تمام بردارهای بازنمایی دادگان تعلیم و آزمون گفتار تمیز و نوفه‌ای مطابق این روابط بازگشتی و به تعداد دور لازم اصلاح می‌شوند.

برای بهره‌گیری مطلوب از این بردارهای بازنمایی اصلاح شده، لازم است آنها را به یک مدل جدید بازشناسی تعلیم دهیم تا بر اساس توزیع جدید بازنمایی‌ها، دوباره خوشبندی مناسب و بازشناسی بهتر صورت گیرد. انتظار است که تعلیم بازنمایی‌های اصلاح شده به یک مدل بازشناسی جدید، منجر به بازشناسی بالاتری نسبت به مدل تعلیم داده شده بر روی دادگان اصلی شود.

مزیت این روش نسبت به روش‌های مشابه که تنها با تخمین خطی اعوجاج‌ها تلاش برای بهبود بازنمایی‌های گفتار می‌کنند در این است که اول این که شبکه دوسویه با استفاده از تخمین غیرخطی سعی در کاهش تنوعات نامطلوب گفتار داشته و بازنمایی‌ها را در جهت افزایش محتوای آوایی آنها بهبود می‌بخشد. دوم این که روش شبکه دوسویه یک روش اصلاحی وابسته به آواست، یعنی به عبارت دیگر، بردارهای بازنمایی گفتار در جهت بهبود صحت بازشناسی آوا اصلاح شده‌اند (Vali et al., 2006b).

۴- طراحی سامانه بازشناسی گفتار

۱-۴- دادگان گفتار

دادگان گفتاری استفاده شده در این تحقیق بخسی از دادگان فارس‌داد است که اولین دادگان گفتاری زبان فارسی می‌باشد. این مجموعه دادگان شامل دو جمله

سطح از بین رفته است. بنابراین برای بهبود و اصلاح بردارهای بازنمایی ورودی بهترین اطلاعات در این سطح شبکه قرار گرفته است (Vali et al., 2006a). در مسیر بازگشتی شبکه BNN از روی این دانش، مؤلفه‌های اصلاحی برای جمع شدن با هر یک از مؤلفه‌های بردارهای بازنمایی ورودی توسط یکتابع غیرخطی تخمین زده می‌شوند. لازم به ذکر است که اگرچه این مؤلفه‌های اصلاحی با مؤلفه‌های بردارهای بازنمایی ورودی جمع می‌شوند، اما به دلیل اینکه ماهیت ویژگی‌ها، لگاریتم انرژی طیف گفتار در باندهای فرکانسی مختلف هستند، بنابراین عبارت جمعی فوق معادل ضرب هر یک از انرژی‌های باندهای فرکانسی در یک دوره جبران‌ساز ناشی از تخمین مسیر بازگشتی شبکه است.

تعلیم بردارهای بازنمایی نوفه‌ای و تمیز به صورت همزمان هم به بخش جلوسوی شبکه و هم به بخش بازگشتی آن کمک بیشتری به اصلاح بردارهای بازنمایی نوفه‌ای و تمیز در جهت بهبود بازشناسی می‌کند (Vali et al., 2006b). بنابراین با توجه به اینکه دادگان تعلیم بخش جلوسو و بازگشتی شبکه یکسان هستند بخش جلوسو و بازگشتی شبکه همزمان با هم تعلیم داده می‌شوند. بدین ترتیب برای هر دور تعلیم ضمن ساخته شدن اطلاعات مفید بازشناسی در لایه مخفی بخش جلوسوی شبکه می‌توان از روی دانش نهفته در این لایه برای اصلاح بردارهای بازنمایی ورودی شبکه در دور بعدی تعلیم استفاده کرد. گفتنی است که در این شیوه در دور اول تعلیم، فقط بخش جلوسوی شبکه برای بردارهای بازنمایی گفتار تمیز و نوفه‌ای به صورت همزمان تعلیم داده می‌شود. در این مرحله وزن‌های بازگشتی شبکه هیچ نقشی ندارند. بنابراین شبکه همان شبکه MLP ساده است. از دور دوم به بعد، بخش بازگشتی شبکه همراه با بخش جلوسوی شبکه در فرایند تعلیم شرکت داده می‌شود. پایداری این شبکه از دو طریق استنباط می‌گردد:

الف: هم‌گرا شدن تعلیم شبکه در جهت افزایش بازشناسی آواهای گفتار.

ب: پس از تعلیم شبکه هر بردار بازنمایی اولیه که به ورودی شبکه داده شود، پس از چند دور چرخش در شبکه مذکور به یک مقدار به طور تقریبی ثابت در رابطه (۷) هم‌گرا شده و در عین حال صحت بازشناسی در طی این دور زدن به صورت صعودی افزایش پیدا کرده تا به یک مقدار به طور تقریبی ثابت برسد.



گرفته شده است. در این روش بازنمایی، فواصل فرکانسی مابین فیلترها و پهنای باند آنها در مقیاس مل تنظیم می‌شوند که به طور تقریبی تا فرکانس یک کیلوهرتز به صورت خطی و بالاتر از این فرکانس به صورت لگاریتمی می‌باشد. در این تحقیق نیز به منظور ایجاد تشابه کامل بانک فیلتر با آنچه در استخراج پارامترهای MFCC رایج است، از مقیاس مل استفاده شده است (So and Paliwal, 2005).

در سامانه‌های بازناسی مقاوم گفتار چنانچه برای جبران‌سازی اثر نوفه جمعی یا کانال از روی بردارهای بازنمایی گفتار، لازم باشد اثر تخریب بر روی هر یک از پارامترها به صورت مجزا شناسایی و بر طرف گردد پارامترهای LFBE بر MFCC ترجیح پیدا می‌کند (Choi, 2005). واضح است که پس از بهبود این نوع بازنمایی چنانچه به بردارهای بازنمایی نوع MFCC نیاز باشد می‌توان با اعمال یک تبدیل DCT بر روی آنها، به پارامترهای MFCC اصلاح شده دست پیدا کرد. و نیز به دلیل همبستگی بیشتر پارامترهای انرژی فیلترهای بانک (LFBE) نسبت به پارامترهای MFCC، در روش‌هایی که به دنبال کاهش میزان همبستگی بین پارامترهای بازنمایی هستند، بازنمایی MFCC کاربرد پیدا می‌کند که منجر به مقاومتر شدن آنها در مقابل نوفه می‌گردد (Jung, 2004). برای به دست آوردن طیف سیگنال از تبدیل فوریه استفاده می‌شود و چون تبدیل فوریه برای سیگنال‌های ایستان تعریف شده و سیگنال گفتار یک سیگنال غیرایستان و شبدهورهای است، لذا باید از تبدیل STFT استفاده شود (Moulines and Verhelst, 1995). به طور معمول طول پنجره و فریم انتخابی جهت تبدیل فوریه زمان کوتاه بین ۱۰ تا ۳۰ میلی ثانیه اختیار می‌شود تا اطمینان حاصل شود که در طول یک فریم سیگنال به طور تقریبی حالت ایستان دارد. با توجه به اینکه فرکانس نمونه برداری سیگنال‌های مورد استفاده ۱۶۰۰۰ هرتز می‌باشد، طول هر پنجره زمانی و گام پیش‌روی، به ترتیب ۲۵۶ و ۱۲۸ نمونه انتخاب شده است که از نظر زمانی شانزده میلی ثانیه می‌باشد و همپوشانی فریم‌ها پنجاه درصد است. پنجره زمانی انتخاب شده از نوع همینگ است. 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 تمیز و نوفه‌ای شامل ۳۹ مؤلفه است.

مشترک بیان شده توسط ۲۰۰ گوینده می‌باشد. دو جمله‌ای که برای این تحقیق در نظر گرفته شدند، از یک جهت بین تمامی گویندگان مشترک بوده و از جهت دیگر به طور تقریبی تمامی واژه‌های زبان فارسی را در بر می‌گیرند (Bijankhan et al., 1994). بنابراین یک سیستم بازناسی گفتار پیوسته مستقل از گوینده ولی وابسته به متن خواهد بود. علت این انتخاب به هدف ما برمی‌گردد که در حقیقت پیداکردن روش‌های بهبود بردارهای بازنمایی با استفاده از قابلیت‌های شبکه‌های عصبی است که وابسته به متن بودن دادگان به قدرت تعیین نتایج این تحقیق لطمه‌ای وارد نمی‌کند. بدین ترتیب از این مجموعه دادگان، ۷۵٪ برای تعلیم شبکه‌ها و ۲۵٪ دیگر آن برای آزمون شبکه‌ها تخصیص داده شده است که گویندگان در دادگان تعلیم و آزمون به طور کامل متفاوت هستند. دادگان نوفه‌ای مورد استفاده در این تحقیق، حاصل اضافه کردن نوفه جمعی سفید گوسی^۱ با نسبت سیگنال به نوفه‌های مختلف (0dB, 5dB, 10dB, 15dB, 20dB) بر روی مجموعه دادگان تمیز فارسی دات، به صورت دستی است. نرخ نمونه برداری سیگنال‌ها برابر ۱۶ کیلوهرتز می‌باشد.

۲-۴- استخراج بازنمایی و LFBE

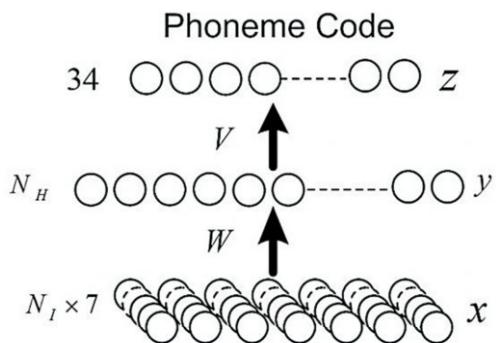
یکی از رایج‌ترین و مناسب‌ترین روش‌های استخراج پارامترهای بازنمایی، ویژگی‌های مبتنی بر طیف سیگنال و حوزه فرکانسی است. در این روش‌ها به کمک بانک فیلترها، بردار ویژگی مربوط به هر فریم به دست می‌آید. بدین صورت که برای هر باند فرکانسی مشخص یک فیلتر قرار داده می‌شود و انرژی (یا دامنه) خروجی هر فیلتر محاسبه شده به عنوان یک پارامتر در نظر گرفته می‌شود. تعداد این فیلترهای میانگذر در سامانه‌های مختلف بازناسی گفتار متفاوت است، ولی به طور معمول بین ۶ الی ۲۰ فیلتر استفاده می‌شود (Chengalvarayan, 2001). افزایش فیلترها به طور معمول موجب بهبود کیفیت بازناسی می‌شود؛ ولی در صورت کاهش پهنای باند فیلترها از فرکانس پایه صدای گوینده، کیفیت بازناسی افت می‌کند (Chengalvarayan, 2001). تنظیم فواصل بین فیلترها می‌تواند به صورت خطی، لگاریتمی و یا به گونه غیرخطی دیگری (مل، بارک) باشد.

پارامترهای لگاریتم انرژی بانک فیلتر^۲ (LFBE) جزو چند روش متداولی است که از دستگاه شنوایی انسان الهام

^۱- White Gaussian Additive Noise

^۲- Logarithm of Filter Bank Energies (LFBE)

پس انتشار خطا با معیار همگرایی کاهش شیب منحنی خطای تعلیم از یک مقدار آستانه است.



(شکل ۳): مدل بازناسی مرجع مبتنی بر شبکه عصبی TDNN

۴-۳- ساختار مدل بازناسی

یکی از بخش‌های اصلی سامانه‌های بازناسی گفتار پس از استخراج بردارهای بازنمایی مناسب از گفتار، مدل بازناسی است. در یک تقسیم‌بندی کلی مدل بازناسی می‌تواند آماری باشد که به طور عمومی مبتنی بر مدل مخفی مارکوف (HMM) است و یا یک مدل هوش مصنوعی باشد که به طور عمومی مبتنی بر شبکه عصبی MLP هستند و یا یک مدل ترکیبی MLP/HMM باشد. از آنجایی که روش‌های ارائه شده برای بازناسی مقاوم گفتار در این مقاله بر مبنای اصلاح بردارهای بازنمایی گفتار مبتنی بر روش کواریانس و شبکه عصبی دوسویه در دادگان مفقود است لازم است یک مدل بازناسی مرجع بر روی بردارهای بازنمایی گفتار تمیز تعلیم داده شود. سپس برای بررسی میزان کارایی روش‌های اصلاح بردارهای بازنمایی، مدل‌های بازناسی جدید با ساختار شبیه به این مدل مرجع بر روی بردارهای بازنمایی اصلاح شده تعلیم داده، ارزیابی می‌شوند و نتایج آنها با مدل مرجع مقایسه خواهد گردید.

اساس مدل‌های بازناسی گفتار به کار رفته در این تحقیق بر مبنای نوع خاصی از شبکه عصبی MLP است که با لحاظ کردن چند فریم متالی به عنوان ورودی، شبیه به ساختار شبکه‌های عصبی^۱ TDNN می‌شود (A. Waibel et al., 1989). لذا در این بخش، ساختار بهینه شبکه عصبی MLP برای دادگان گفتاری این تحقیق، شناسایی و طراحی و براساس آن مدل بازناسی بر روی بردارهای بازنمایی گفتار تمیز تعلیم داده می‌شود تا در ادامه به عنوان مرجع مقایسه از آن استفاده شود.

(شکل ۳) ساختار نورونی یک شبکه عصبی MLP را با یک لایه پنهان به صورت $N_I \times 7 \rightarrow N_H \rightarrow 34$ نشان می‌دهد. در ورودی، بردار بازنمایی فریم جاری به همراه بردارهای بازنمایی سه فریم مجاور چپ و راست (در مجموع هفت فریم) قرار می‌گیرند، درنتیجه ورودی شبکه $N_I = 7$ نورون خواهد شد. خروجی شبکه نیز به تعداد کدهای آوایی موجود در دادگان تعلیم فارسی یعنی ۳۴ نورون در نظر گرفته شده است.

شبکه ذکر شده بر روی ویژگی‌های MFCC استخراج شده از ویژگی‌های LFBE، تعلیم داده می‌شود که تعداد نورون‌های لایه مخفی آن یکصد نورون انتخاب شده است. این تعداد نورون بر اساس حجم دادگان تعلیمی شبکه، به صورت بهینه انتخاب شده است (Blumer et al., 1989). مدل بازناسی، بر روی بردارهای بازنمایی دادگان گفتاری تمیز تا همگرایی کامل تعلیم داده می‌شود. الگوریتم تعلیم

۵- شرح آزمایش‌ها

۵-۱- ارزیابی مدل مرجع بر روی دادگان تست تمیز و نویزی

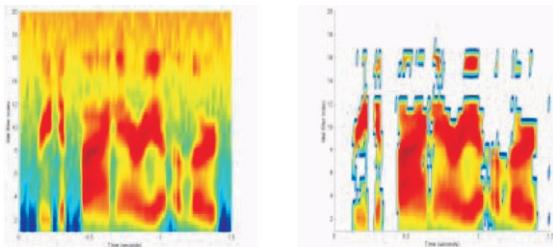
از آنجایی که در این تحقیق چندین بار بردارهای بازنمایی گفتار تمیز و نویزهای با چندین روش مختلف اصلاح می‌شوند برای ارزیابی میزان موفقیت روش‌های پیشنهادی در اصلاح بردارهای بازنمایی، لازم است مدل‌های بازناسی مجازایی شبیه به ساختار مدل مرجع برای ارزیابی روش‌های مختلف اصلاح، بر روی دادگان گفتار تمیز اصلاحی آن روش، تعلیم داده شده و بر روی دادگان آزمون نویزهای اصلاحی نظیر آن نیز ارزیابی شوند. در این تحقیق همان‌طور که اشاره شد، دادگان نویزی شامل گفتار تمیزی است که با SNR های مختلف (0dB, 5dB, 10dB, 15dB & 20dB) توسط نویزهای سفید گوسی جمعی تخریب شده است. (جدول ۱) صحبت بازناسی مدل مرجع برای دادگان آزمون گفتار تمیز و نویزی در مدل بازناسی مرجع بر حسب درصد صحت فریم

(جدول ۱): صحبت بازناسی برای دادگان آزمون گفتار تمیز و نویزی در مدل بازناسی مرجع بر حسب درصد صحت فریم

SNR in dB	Frame Accuracy (%)
۳۰ dB	۸۳/۴
۲۰ dB	۷۸/۳
۱۵ dB	۷۲/۷
۱۰ dB	۶۴/۶
۵ dB	۵۳/۳
۰ dB	۴۰/۳

^۱ Time Delay Neural Network (TDNN)





(شکل ۳): سمت چپ، اسپکتروگرام یک گفتار نویه‌ای و شکل سمت راست، ماسک اسپکتروگرام همان گفتار نویه‌ای زمانی که تمامی مؤلفه‌های با SNR پایین تر از حد آستانه ۳dB از اسپکتروگرام پاک شده‌اند، نمایش می‌دهد.

در ارزیابی اولیه مشاهده شد که روش مبتنی بر کواریانس عملکرد چشم‌گیری در بازناسی نواحی مفقود داشته است. با این اوصاف انتظار می‌رفت که بهبود قبل ملاحظه‌ای نیز در بازناسی گفتارهای تخریب‌شده توسط نویز با SNRهای مختلف داشته باشد، لذا در (شکل ۵) نتایج بازناسی روی بردارهای بازنمایی MFCC اصلاح شده توسط روش مبتنی بر کواریانس در روش ماسک از پیش‌دانسته ارائه شده است. مدل بازناسی، همان شبکه مرجع است که با دادگان تمیز تعلیم یافته و با دادگان اصلاح شده در SNRهای مختلف ارزیابی صورت گرفته است. نتایج به دست آمده، قابلیت روش را در بازناسی نواحی تخریب شده توسط نویز، به‌خوبی نشان می‌دهد. این بهبود (در صحت بازناسی) حداقل ۲/۳ درصد برای بازناسی گفتارهای نویه‌ای با SNR=20dB و حداقل ۱۵/۲ درصد برای بازناسی گفتارهای تخریب‌شده توسط نویز با SNR=0dB نمود پیدا کرد.

در عمل پیاده‌سازی چنین سامانه‌ای میسر نیست چرا که هیچ دانشی از میزان نویه تخریبی و نواحی تخریب شده توسط آن در دسترس نیست و باید سامانه، طوری طراحی شود که به صورت خودکار قادر به شناسایی مؤلفه‌های مفقود اسپکتروگرام باشد.

حال باید دید زمانی که هیچ دانش اولیه‌ای از نویه داشته باشیم، روش مبتنی بر کواریانس تا چه حد قابلیت جبران اثر نویه را خواهد داشت. بعد از اطمینان از عملکرد سامانه بازناسی، در ادامه از تخمین نویه با استفاده از روش سری تیلور برداری که شرح آن در بخش (۱-۲) رفت، برای شناسایی مؤلفه‌های مفقود استفاده شد و بازناسی مؤلفه‌های

۲-۵- اصلاح بردارهای بازنمایی لگاریتم طیفی با استفاده از روش دادگان مفقود

در این بخش روش بازناسی مبتنی بر کواریانس برای جبران اثر نویه جمعی ارزیابی خواهد شد. این ارزیابی طی دو مرحله صورت خواهد گرفت. در مرحله اول صرفاً ارزیابی صحت بازناسی مؤلفه‌ها مورد نظر است. اما مرحله دوم ارزیابی بر روی صحت بازناسی است که ابتدا شامل شناسایی مؤلفه‌های مفقود و سپس بازناسی آنها خواهد بود. بردارهای بازنمایی گفتار، LFBهای ۲۰ مؤلفه‌ای هستند که پس از بازناسایی شدن مؤلفه‌های مفقود آنها، با اعمال تبدیل DCT به مؤلفه‌های MFCC تبدیل می‌شوند.

در مرحله اول برای ارزیابی صحت بازناسی، نیازی به شناسایی مؤلفه‌های مفقود نیست؛ چرا که مؤلفه‌های مفقود با دانش قلی از نویه تعیین می‌شوند که به آن ارزیابی از روی ماسک از پیش‌دانسته^۱ می‌گوییم. در روش ماسک از پیش‌دانسته، اسپکتروگرامی که مؤلفه‌های مفقود آن از قبل شناسایی و پاک شده‌اند، برای ارزیابی صحت بازناسی استفاده می‌شود و این تنها تفاوت بین روش ماسک از پیش‌دانسته و روش ارزیابی کلی است. این روش می‌تواند یک ارزیابی اولیه مفیدی از عملکرد روش بازناسی مبتنی بر کواریانس محاسبه شود.

هنگامی که سیگنال گفتار توسط یک نویه جمعی غیر همبسته تخریب می‌شود، طیف توان سیگنال حاصل از جمع طیف توان گفتار تمیز با طیف توان نویه طبق رابطه (۱۱) به دست می‌آید. زمانی که طیف توان نویه جمعی N_p را در اختیار داشته باشیم، از این طریق می‌توان طیف توان سیگنال تمیز X_p را از طیف توان سیگنال نویه‌ای Y_p به دست آوریم و سپس می‌توان SNR را به صورت محلی از رابطه (۱۲) به دست آورد:

$$Y_p(m, k) = X_p(m, k) + N_p(m, k) \quad (11)$$

$$SNR(m, k) = 10 \log_{10} \frac{X_p(m, k)}{N_p(m, k)} \quad (12)$$

که (۱۱) و $N_p(m, k)$ و $X_p(m, k)$ به ترتیب طیف توان گفتار تمیز و نویه را در باند فرکانسی k ام از فریم m اسپکتروگرام نشان می‌دهند. حال می‌توان تمامی مؤلفه‌های اسپکتروگرام را که از یک حد آستانه ثابت، SNR پایین‌تری را داشته باشند، به عنوان مؤلفه‌های مفقود در نظر گرفت و از بردارهای بازنمایی LFBهای حذف کرد. چون مقدار ثابتی برای حد آستانه وجود ندارد، ما در این تحقیق، حد آستانه‌های متفاوتی را مورد ارزیابی قرار دادیم و بهترین حد آستانه را

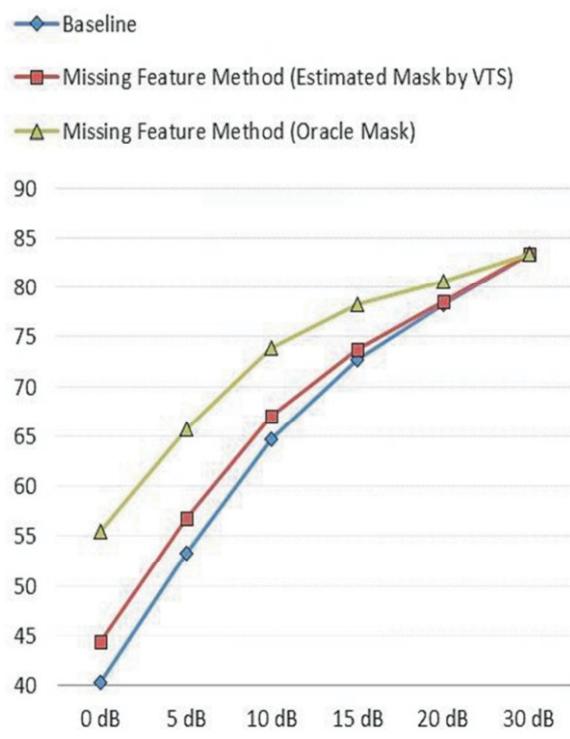
^۱ Oracle Mask

می‌گیرد. نحوه تعلیم شبکه دوسویه، طبق روابط آمده در (بخش ۱-۳) است. یعنی وزن‌های بخش جلوسو و بازگشتی شبکه، به صورت همزمان در هر دور تعلیم، اصلاح می‌شوند. در مرحله آزمون شبکه، دادگان آزمون مطابق روابط (۷) تا (۱۰) برای $n = 1, 2, \dots, 6$ دور در شبکه چرخانده شده و هر بار صحت بازشناسی در خروجی شبکه دوسویه برای آنها به دست آمده است. $n = 1$ به منزله این است که فقط بخش جلوسوی شبکه دوسویه در بازشناسی دخالت داده شده است. پس از اولین دور مقادیر خروجی نورون‌های لایه مخفی بخش جلوسوی شبکه (لایه z) برای تمام دادگان آزمون مشخص شده و از آنها برای اصلاح ورودی و بازشناسی آواها برای شش دور محاسبه می‌شوند تا هم‌گرایی الگوریتم گردش در شبکه عصبی دوسویه ارزیابی شود. نمودارهای (شکل ۶) نشان می‌دهند که در هر دور محاسبه، اصلاح بردارهای بازنمایی ورودی، در جهت افزایش صحت بازشناسی صورت گرفته است. این مسئله هم برای دادگان تمیز و هم برای دادگان نوشهای حاصل می‌شود. از تکرار $n = 4$ به بعد نتایج به طور تقریبی ثابت و پایدار شده‌اند. لذا شبکه عصبی دوسویه در فرآیند اصلاح بردارهای بازنمایی آزمون، طی چند دور چرخش در جهت افزایش صحت بازشناسی و تثبیت پیش‌رفته و می‌توان اطمینان پیدا کرد که فرآیند اصلاح بردارهای بازنمایی در این شبکه، بعد از چهار دور به طور تقریبی متوقف شده است و بازنمایی‌های جدید تولید شده در رابطه (۷) را بتوان بعداً در مدل‌های بازشناسی استفاده کرد. تا اینجا هر یک از شبکه‌های دوسویه پنج گانه بالا بر روی یک دسته دادگان با SNR مشخص تعلیم داده شده است و اصلاح نیز بر روی دادگان آزمون نظیر همان دادگان تعلیم صورت گرفته است. اما لازم است روشی ارائه شود که هر نوع داده را برای SNR دلخواه بتواند بدون اینکه نیاز باشد همه دادگان با انواع SNR به یک شبکه عصبی دوسویه تعلیم داده شود را در یک فرآیند پیشنهادی اصلاح کند. برای این منظور پیشنهاد زیر مطرح می‌گردد:

ابتدا بردارهای بازنمایی LFBE اصلاح شده برای تمام دادگان تعلیم و آزمون مطابق روابط (۷) تا (۱۰) برای همه SNRها و برای هر پنج شبکه دوسویه یادشده به دست می‌آیند. با اعمال تبدیل DCT بر روی بردارهای بیست مؤلفه‌ای اصلاحی LFBE، بردارهای ۱۳ MFCC تایی اصلاحی، حاصل می‌شوند که با قرار دادن مشتقات اول و دوم کنار آنها، بردارهای بازنمایی ۳۹ مؤلفه‌ای اصلاحی MFCC به دست می‌آیند.

شناسایی شده، توسط همان شیوه مبتنی بر کواریانس انجام پذیرفت. در همان نسودار (شکل ۵) ارزیابی کلی روش بازسازی مبتنی بر کواریانس را بر روی ۳۹ ویژگی MFCC زمانی که از تخمین نوافه برای شناسایی مؤلفه‌های مفقود استفاده شده، نشان داده شده است.

همان‌طور که (شکل ۵) نشان می‌دهد زمانی که نواحی تخریب شده اسپکتروگرام از قبل شناخته شده باشند، نتایج صحت بازشناسی به دست آمده از روش‌های بازسازی، خیلی بالاتر از زمانی است که از روش مبتنی بر کواریانس در حوزه ویژگی‌های مفقود استفاده شود؛ با این وجود، صحت بازشناسی ۴۴/۵ درصدی و بهبود ۴/۲ درصدی بدست آمده برای بازسازی گفتارهای تخریب شده توسط نوافه با SNR=0dB نشان می‌دهد که عملکرد روش کواریانس در شناسایی و بازسازی نواحی مفقود و بالا بردن صحت بازشناسی مؤثر بوده و می‌توان از این روش به عنوان یکی از روش‌های موفق در جبران‌سازی اثر نوافه استفاده کرد.



(شکل ۴): صحت بازشناسی به دست آمده از اصلاح بردارهای بازنمایی MFCC با استفاده از روش دادگان مفقود با ماسک از پیش‌دانسته و تخمین ماسک

۵-۳- اصلاح بردارهای بازنمایی با شبکه عصبی دوسویه

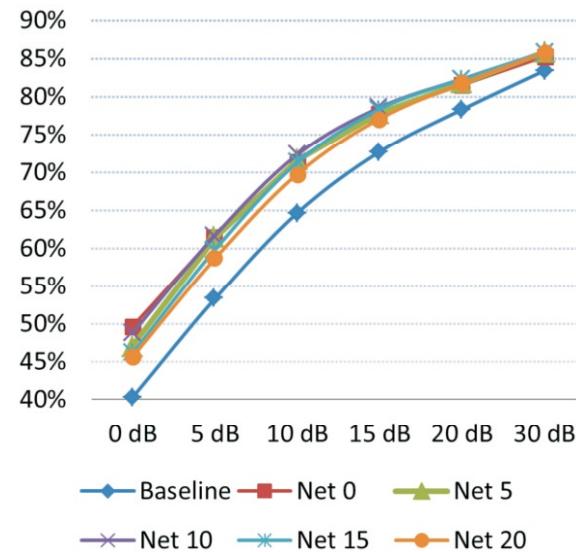
در این بخش، توانایی شبکه عصبی دوسویه برای اصلاح غیرخطی بردارهای بازنمایی LFBE مورد بررسی قرار



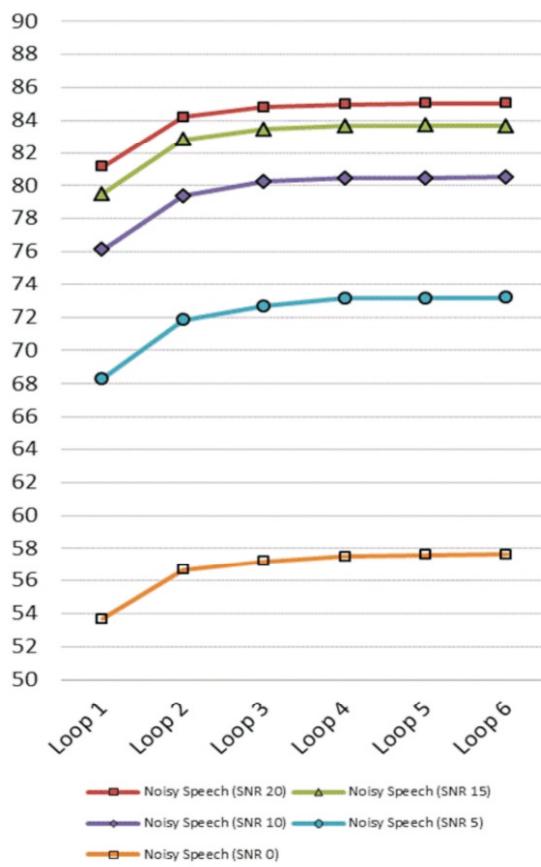
بازنایمی‌های نوفه‌ای در تمام SNR ها داشته‌اند که البته این بهبود در ۵ Net، یعنی زمانی که از گفتار تمیز و نوفه‌ای با نسبت سیگنال به نویه ۵dB به طور هم‌زمان در تعلیم شبکه عصبی دوسویه استفاده شده، بیشتر نمود پیدا کرده است که منجر به افزایش ۱/۷ درصدی در بازناسی سیگنال تمیز و همچنین ۲/۶ الی هشت درصدی در SNR=20dB بازشناختی گفتارهای تخریب شده توسط نوفه با SNR=0dB شده است. با وجود بهبود قابل ملاحظه در ۵ Net، (جدول ۲) به طور واضح نشان می‌دهد که در برخی از شبکه‌های دوسویه، در برخی از SNRها، بهبودی در صحبت بازناسی حاصل نشده است و یا این بهبود چندان رضایت بخش نیست.

لذا برای ارتقای نتایج، در ادامه مقادیر پویای پارامترهای بازنایمی LFBE (مشتقات مرتبه اول و دوم) در کنار ایستای آنها به کار گرفته می‌شود تا اصلاح مشتقات بازنایمی نیز توسط شبکه دوسویه صورت گیرد.

برای ارزیابی، شبکه عصبی دوسویه، همانند قبل با ساختار نشان داده شده در (شکل ۲) ولی این بار بر روی بردارهای بازنایمی صفت‌تایی تمیز و نوفه‌ای به صورت هم‌زمان تعلیم داده می‌شود. ساختار نورونی شبکه عصبی دوسویه به صورت $N_r = 50$, $N_H = 100$, $N_I = 7 \cdot 60$ و $N_0 = 34$ می‌باشد. هم‌گرایی مطلوب شبکه نیز همانند قبل پس از پنجاه دور تعلیم حاصل می‌شود. همان‌طور که در قبل گفته شد برای تعیین اینکه کدام مجموعه گفتار نوفه‌ای با چه نسبت سیگنال به نوفه‌ای برای تعلیم به شبکه دوسویه همراه با گفتار تمیز مناسب است، لازم است که برای هر دادگان گفتاری نوفه‌ای با نسبت سیگنال به نویز معین یک شبکه عصبی دوسویه همراه با دادگان تمیز تعلیم داده شود.



(شکل ۶): صحبت بازناسی به دست آمده از اصلاح به روش شبکه عصبی دوسویه برای بردارهای بازنایمی MFCC استخراج شده از بردارهای بازنایمی صفت‌تایی LFBE



(شکل ۵): میزان افزایش صحبت بازناسی آواهای گفتار آزمون در اثر شش دور محاسبه در شبکه دوسویه تعلیم داده شده بر روی گفتار تمیز و نوفه‌ای برای هر SNR

حال برای تعیین اینکه کدام یک از پنج شبکه دوسویه برای استفاده جهت اصلاح کلیه بردارهای بازنایمی با SNR های مختلف موفق‌تر عمل کرده است، پنج مدل بازناسی مبتنی بر شبکه عصبی MLP شبیه به مدل مرجع (در قبل بر روی بردارهای بازنایمی تمیز تعلیم داده شده بود) بر روی بازنایمی‌های MFCC تمیز اصلاحی استخراج شده از هر یک از پنج شبکه دوسویه، تعلیم داده می‌شود. (جدول ۲) نتایج بازناسی گفتار بر روی دادگان آزمون اصلاحی با انواع SNR را در پنج مدل بازناسی مذکور نشان می‌دهد.

(جدول ۲): صحبت بازناسی به دست آمده از اصلاح بردارهای بازنایمی به روش شبکه عصبی دوسویه برای بردارهای بازنایمی MFCC

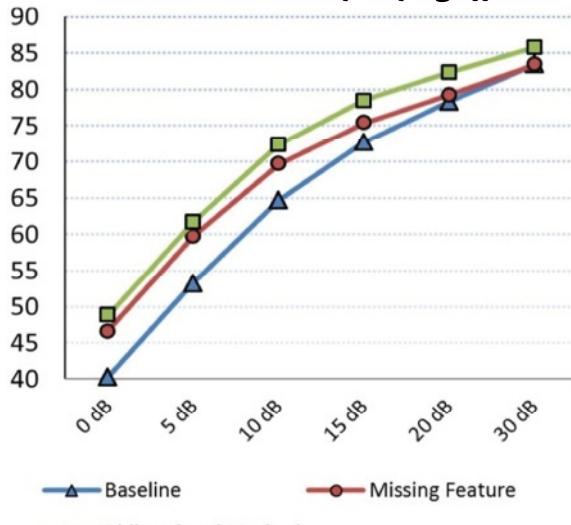
SNR in dB	Baseline	Net 0	Net 5	Net 10	Net 15	Net 20
۳۰ dB	۸۳/۴	۸۴/۵	۸۵/۱	۸۴/۶	۸۴/۵	۸۵/۰
۲۰ dB	۷۸/۳	۸۰/۲	۸۱/۳	۸۰/۵	۸۰/۰	۸۰/۰
۱۵ dB	۷۲/۲	۷۶/۲	۷۷/۴	۷۵/۸	۷۵/۵	۷۴/۹
۱۰ dB	۶۴/۶	۶۹/۹	۷۰/۹	۶۸/۹	۶۸/۰	۶۶/۳
۵ dB	۵۳/۳	۶۰/۲	۶۱/۲	۵۸/۴	۵۶/۳	۵۳/۹
۰ dB	۴۰/۳	۴۷/۶	۴۸/۳	۴۵/۶	۴۳/۱	۴۰/۲

همان‌طور که در (جدول ۲) مشاهده می‌شود، هر پنج شبکه عصبی دوسویه پیشرفت قابل قبولی در بهبود

نوفه نزدیک به 10 dB در کنار گفتار تمیز برای تعلیم به شبکه دوسویه مناسب است.

بدین ترتیب نحوه استفاده بهینه از شبکه دوسویه برای اصلاح بردارهای بازنمایی، به صورت زیر خلاصه می‌شود:

- ابتدا از کلیه بردارهای بازنمایی گفتار تمیز و نوفه‌ای در سطوح مختلف SNR، بردارهای بازنمایی بیست مؤلفه‌ای LFBE استخراج می‌شوند.
- با اضافه کردن مشتقات اول و دوم بازنمایی در کنار مؤلفه‌های ایستا، بازنمایی‌های شصت مؤلفه‌ای ساخته می‌شوند.
- یک شبکه عصبی دوسویه بر روی دادگان تمیز و نوفه‌ای با SNR در محدوده صفر تا ده دسی بل تعلیم داده می‌شود.
- کلیه دادگان تمیز و نوفه‌ای در سطوح مختلف SNR در طی چهار دور گردش در شبکه دوسویه ذکر شده اصلاح می‌شوند.
- دادگان تمیز اصلاح شده از بالا برای تعلیم به هر مدل بازناسنی گفتار قبل استفاده‌اند که بر روی تمام دادگان نوفه‌ای اصلاح شده ذکر شده نیز قابل ارزیابی خواهد بود.



(شکل ۷): صحّت بازناسنی به دست آمده از روش‌های دادگان مفقود و شبکه عصبی دوسویه

۶- جمع بندی

هدف از این تحقیق، ارائه یک راهکار جدید در بازناسنی مقاوم گفتار بوده است که تا جای ممکن روش ارائه شده برای جرمان‌سازی بیشتر تنوعات گفتار، جواب‌گو باشد. شیوه کلی مواجهه با بازناسنی مقاوم گفتار در بهبود و اصلاح بازنمایی‌های گفتار قبل از ورود به مدل بازناسنی بوده است. زیرا همان طور که می‌دانیم بهترین روش‌های دسته‌بندی الگو نیز در مقابل الگوهای تخریب‌شده که اطلاعات مفید و

پس از تعلیم پنج شبکه دوسویه ($x = 0, 5, 10, 15, 20$) و اصلاح کلیه بردارهای بازنمایی صفت مؤلفه‌ای LFBE در کلیه SNR ها و سپس تبدیل بردارهای LFBE به MFCC و تعلیم مجدد پنج مدل بازناسنی مبتنی بر ساختار مرجع بر روی کلیه دادگان اصلاح شده (کلیه SNR ها) در هر یک از Net x ها، نتایج صحّت بازناسنی برای دادگان آزمون اصلاحی در SNR های مختلف به دست آمد که در نمودار (شکل ۷) گزارش شده است. همان‌طور که مشاهده می‌شود در اصلاح بردارهای بازنمایی، استفاده از مقادیر دینامیک پارامترها در کنار ایستای آنها نقش مؤثّری در ارتقای صحّت بازناسنی داشته است و هر پنج شبکه به صورت محسوسی نتایج را برای کلیه SNR ها ارتقا داده‌اند چراکه هر چه شبکه دوسویه در شناسایی آواها عملکرد بهتری داشته باشد دانش آوایی نهفته در لایه پنهان بخش جلوسوی شبکه نیز قوی‌تر بوده و اصلاح ویژگی‌های آواها با دقت بیشتری صورت می‌پذیرد. از نمودار (شکل ۷) واضح است که تقریبا هر کدام از پنج شبکه دوسویه فوق در اصلاح بردارهای بازنمایی مربوط به SNR نظری خود موفق‌تر از سایر شبکه‌ها عمل کرده است زیرا طبیعی است که دادگان تعلیمی به یک شبکه دوسویه بهتر از سایر دادگانی که در زمان تعلیم مشاهده نشده‌اند اصلاح خواهد شد.

بهترین نتایج مربوط به Net 10 است که بهبود ۴۰٪ ۸/۵ درصدی در صحّت بازناسنی گفتارهای نوفه‌ای به ترتیب برای $SNR=0dB$ و $SNR=20dB$ داشته است. بدین ترتیب نتایج نمودارهای (شکل ۷) هم عملکرد چشم‌گیر شبکه‌های عصبی دوسویه در بازسازی مؤلفه‌های مفقود را نشان می‌دهد و هم بیان می‌کند زمانی که شبکه به‌وسیله بردارهای بازنمایی شصت تایی LFBE تعلیم داده شود میزان بهبود بیشتر و قابل اطمینان تر خواهد بود. برای مقایسه بهتر، دو روش پیشنهادی اصلاح ویژگی‌های ایستا و پویا با استفاده از شبکه دوسویه، در (جدول ۳) میزان میانگین بهبود در صحّت بازناسنی برای کلیه SNR ها آورده شده است.

(جدول ۳): میزان میانگین بهبود صحّت بازناسنی در دو روش پیشنهادی استفاده از شبکه عصبی دوسویه

AVERAGE OF IMPROVEMENT IN RECOGNITION ACCURACY	NET 0	NET 5	NET 10	NET 15	NET 20
STATIC FEATURES: LFBE = 20	۴/۱۱	۵/۰۴	۳/۳۱	۲/۲۴	۱/۰۶
STATIC + DYNAMIC FEATURES : LFBE = 60	۵/۴۹	۵/۲۴	۵/۹۲	۴/۹۶	۴/۱۲

با توجه به (جدول ۳) می‌توان ادعا کرد زمانی که از پویایی بردارهای بازنمایی در کنار ایستای آنها (بردارهای بازنمایی شصت تایی LFBE) برای تعلیم شبکه عصبی دوسویه استفاده شود، به کارگیری گفتارهای نوفه‌ای با سطح

- Acero A. (1993) Acoustic and Environmental Robustness in Automatic Speech Recognition. illustrated ed. Springer Publishers, Boston.
- Bijankhan M., Sheikhzadegan J., Roohani M., Samareh Y., Lucas C., Tebyani M. (1994) FARSDAT-The speech database of Farsi spoken language, International Conference on Speech Sciences & Technology. pp. 826-830.
- Blumer A., Ehrenfeucht A., Haussler D., Warmuth M.K. (1989) Learnability and the Vapnik-Chervonenkis dimension. Journal of the ACM (JACM) 36: 929-965.
- Boll S.F. (1979) Suppression of Acoustic Noise in Speech Using Spectral Subtraction. IEEE Transactions on Acoustics, Speech and Signal Processing 27:113-120. DOI: 10.1109/TASLSP.1979.1163209.
- Bourlard H., Dupont S. (1996) A new ASR approach based on independent processing and recombination of partial frequency bands, IEEE International Conference on Spoken Language Proceedings, IEEE, Philadelphia, PA , USA. pp. 426-429.
- Chengalvarayan R. (2001) Evaluation of front-end features and noise compensation methods for robust mandarin speech recognition, European Conference on Speech Communication and Technology: EUROSPEECH, Aalborg, Denmark. pp. 897-900.
- Choi E.H.C. (2005) A Generalized Framework for Compensation of Mel-Filterbank Outputs in Feature Extraction for Robust ASR, Interspeech'2005 - Eurospeech, Lisbon, Portugal. pp. 933-936.
- Cooke M., Green P., Crawford M. (1994) Handling Missing Data in Speech Recognition, International Conference on Spoken Language Processing, Yokohama, Japan. pp. 1555-1558.
- Gales M.J.F., Young S.J. (1993) HMM Recognition in Noise Using Parallel Model Combination, European Conference on Speech Communication and Technology: EUROSPEECH' 93, Berlin, Germany pp. 837-840.
- Hennansky H., Tibrewala S., Pavel M. (1996) Towards ASR on partially corrupted speech, IEEE International Conference on Spoken Language Proceedings, IEEE, Philadelphia, PA , USA. pp. 462-465.
- Jung H.Y. (2004) Filtering of filter-bank energies for robust speech recognition. ETRI journal 26:273-276.
- Leggetter C.J., Woodland P.C. (1994) Speaker adaptation of HMMs using linear regression. Cambridge University, Cambridge, UK, Tech. Rep. CUED/F-IN-FENG/TR.181 181.
- Lippmann R., Carlson B.A. (1997) Using Missing Feature Theory to Actively Select Features for Robust Speech Recognition with Interruptions, Filtering and Noise, European Conference on Speech Communication and Technology: EUROSPEECH, Rhodes, Greece. pp. 37-40.

متمايز گننده را از دست داده باشنند، توانايي مطلوبی خواهند داشت. نقطه قوت روش اصلاح و پرگى مبتنی بر شبکه عصبی دوسویه نسبت به روش و پرگى های مفقود، این است که نیازی به شناسایی مؤلفه های مفقود ندارد و بازسازی را در جهت هرچه شبیه تر شدن تمامی مؤلفه های معتبر یا نامعتبر به مؤلفه های گفتار تمیز صورت می دهد و این یک برتری بسیار چشم گیری است که در این تحقیق حاصل شده است؛ چرا که در عمل، بحث شناسایی مؤلفه های مفقود، که یک بحث چالش برانگیز در تمامی روش های بازناسی مقاوم گفتار مبتنی بر و پرگى های مفقود است و ارتباط مستقیمی با میزان صحبت بازناسی دارد را حذف می کند. ایده اصلی نهفته در تمام روش های ارائه شده در این تحقیق در استفاده از دانش مشترکی است که در بردارهای بازنمایی گفتار تمیز و نوشهای وجود دارد و می توان اطلاعات مفیدی را که در اثر نوشه از بین رفته است از بازنمایی های سالم بیرون کشید و در بهبود بازنمایی های نوشهای به کار گرفت.

بررسی و مقایسه کارایی دو نوع روش جبران اثر نوشه، دادگان مفقود و شبکه های عصبی دوسویه در مدل بازناسی بر روی گفتار نوشهای و تمیز و همچنین کاربرد تکنیک شبکه های عصبی دوسویه در اصلاح بردارهای بازنمایی در جهت زدیک کردن بازنمایی های دسته های آوایی یکسان از دو نوع گفتار نوشهای و تمیز به هم دیگر و دور کردن بازنمایی های دسته های آوایی متفاوت از هم دیگر از نوآوری های این تحقیق به شمار می رفت که شرح آن داده شد.

پیشرفت چشم گیری که در روند بازسازی با استفاده از روش جبران سازی توسط شبکه های عصبی دوسویه در مقایسه با بازسازی مؤلفه های مفقود مبتنی بر روش کواریانس حاصل شد، در نمودار (شکل ۸) به وضوح دیده می شود.

(شکل ۸): به خوبی نشان می دهد که صحبت بازناسی به دست آمده از روش جبران سازی توسط شبکه های عصبی دوسویه در بهترین و بدترین حالت به ترتیب بهبود ۴/۵ و ۴/۳ درصدی نسبت به حالت پایه داشته است که معرف سیگنال نوشه ای بدون هیچ جبران سازی اثر نوشه می باشد. همچنین بهبود ۴/۳ و ۳/۷ درصدی نیز نسبت به روش جبران اثر نوشه مبتنی بر کواریانس در SNR های مختلف به دست آورده است.

۷- مراجع

- A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, K. J. Lang. (1989) Phoneme recognition using time-delay neural networks. IEEE Transactions on Acoustics, Speech and Signal Processing 37:328-339.



حاجت محمدنژاد تحصیلات خود را در مقطع کارشناسی مهندسی مخابرات دانشگاه ارومیه در سال ۱۳۸۵ و در مقطع کارشناسی ارشد مهندسی بیوالکتریک دانشگاه شاهد در سال ۱۳۸۹ به پایان رسانده است. وی از همان سال ۱۳۸۹ در پژوهشگاه مخابرات و الکترونیک نصر در گروه رمز مشغول به فعالیت است. حوزه‌های تحقیقاتی تخصصی وی عبارتند از: پردازش سیگنال، پردازش گفتار، پردازش داده رمز، بازناسی الگو، شبکه‌های عصبی، ماشین بردار پشتیبان.

نشانی رایانماء ایشان عبارت است از
hojat.mohammadnejad@gmail.com



منصور ولی در سال ۱۳۸۵ تحصیلات مقطع دکتری خود را در مهندسی پزشکی (بیوالکتریک) در دانشگاه صنعتی امیرکبیر به اتمام رساند. ایشان مدرک کارشناسی ارشد خود را سال ۱۳۷۸ در مهندسی پزشکی (بیوالکتریک) از دانشگاه صنعتی شریف و مدرک کارشناسی خود را سال ۱۳۷۶ در مهندسی الکترونیک از دانشگاه صنعتی اصفهان دریافت نمود. وی در حال حاضر استادیار گروه مهندسی پزشکی دانشکده برق دانشگاه صنعتی خواجه نصیرالدین طوسی می باشد. زمینه‌های پژوهشی مورد علاقه ایشان پردازش گفتار، پردازش صوت و کاربرد آن در مهندسی پزشکی و شبکه‌های عصبی مصنوعی و زیستی است.

نشانی رایانماء ایشان عبارت است از
mansour.vali@eetd.kntu.ac.ir

Moore B.C.J. (1997) An Introduction to the Psychology of Hearing. Illustrated ed. Academic Press Inc, San Diego.

Moreno P.J. (1996) Speech Recognition in Noisy Environments, Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania.

Moulines E., Verhelst W. (1995) Time-domain and frequency-domain techniques for prosodic modification of speech. Speech coding and synthesis:519-555.

Neumeyer L., Weintraub M. (1994) Probabilistic Optimal Filtering for Robust Speech Recognition, IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, Adelaide, SA , Australia pp. I/417 - I/420

Porter J.E., Boll S.F. (1984) Optimal estimators for spectral estimators of noisy speech, IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE. pp. 53 - 56.

Raj B. (2000) Reconstruction of Incomplete Spectrograms for Robust Speech Recognition, Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania. pp. 219.

Raj B., Stern R.M. (2005) Missing-Feature Approaches in Speech Recognition: Improving recognition accuracy in noise by using partial spectrographic information. Signal Processing Magazine, IEEE 22: 101-116.

Raj B., Parikh V.N., Stern R.M. (1997) The effects of background music on speech recognition accuracy, IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, Munich , Germany. pp. 851-854.

So S., Paliwal K.K. (2005) Improved noise-robustness in distributed speech recognition via perceptually-weighted vector quantisation of filterbank energies, Interspeech'2005 - Eurospeech, Lisbon, Portugal. pp. 941-944.

Vali M., Salehi S.A.S., Karimi K. (2006a) Improvement of Feature Vectors in Clean and Telephone Speech Recognition using Bidirectional Neural Network, Speech Recognition and Intrinsic Variation, Toulouse, France. pp. 41-46.

Vali M., Salehi S.A.S., Karimi K. (2006b) Robust Speech Recognition by Modifying Clean and Telephone Feature Vectors Using Bidirectional Neural Network, International Conference on Spoken Language Processing, Pittsburgh, PA, USA. pp. 2072.

Varga A.P., Moore R.K. (1990) Hidden Markov model decomposition of speech and noise, IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, Albuquerque, NM , USA. pp. 845-848.