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

حجّت محمّدنژاد[1]، منصور ولی[2]
[1] دانشگاه شاهد، دانشکدهٔ فنّی
[2] دانشگاه‌شاهد، دانشکدهٔ فنّی و دانشگاه صنعتی خواجه‌نصیرالدین طوسی، دانشکدهٔ برق و کامپیوتر

چکیده

عملکرد سامانه‌های بازشناسی گفتار زمانی که گفتار توسط نوفه تخریب شده باشد، به‌شدّت کاهش می‌یابد. یکی از روش‌های رایج برای مقاوم‌سازی سامانه‌های بازشناسی گفتار استفاده از روش دادگان مفقود است. در این روش مؤلّفه‌هایی از نمایش زمانی- فرکانسی گفتار (اسپکتروگرام) که نسبت سیگنال به نوفه (SNR) آنها از یک آستانه کم‌تر است، به‌عنوان مؤلّفه‌های نامعتبر یا مفقود برچسب‌دهی می‌شوند. این مؤلّفه‌ها با استفاده از مؤلّفه‌های معتبر و اطّلاعات آماری نسبت به دادگان تمیز، تخمین زده شده و جایگزین می‌شوند.

در این مقاله الگوی ویژگی‌های مفقود با دیدگاهی نو، به‌عنوان مسئلهٔ جبران‌سازی دادگان مطرح می‌شود. بدین‌صورت که با استفاده از شبکهٔ عصبی دوسویه و انجام یک سری پردازش‌های غیرخطی و دوطرفه (جلوسو و برگشتی) از دانش نهفته در مدل، ناشی از یادگیری هم‌زمان گفتار تمیز و نوفهی بهره گرفته، بردارهای بازنمایی گفتار در جهت افزایش صحت بازشناسی آواها بهبود می یابند.

در این روش نیازی به شناسایی مؤلّفه‌های مفقود که یک بحث چالش برانگیز در حوزهٔ بازشناسی مقاوم گفتار مبتنی‌بر دادگان مفقود است، نمی‌باشد؛ بلکه بازسازی در جهت هرچه شبیه‌ترشدن تمامی مؤلّفه‌ها، خواه معتبر باشد، خواه نامعتبر به مؤلّفه‌های گفتار تمیز صورت می‌گیرد و این یک برتری بسیار چشم‌گیری است که در این تحقیق حاصل شده است. نتایج مقایسهٔ این دو روش نشان می‌دهد که با استفاده از روش دادگان مفقود، ۴/۲ درصد بهبود برای صحّت بازشناسی گفتار نوفه‌ای با نسبت سیگنال به نوفه ۰dB حاصل شده درحالی‌که با استفاده از روش مبتنی‌بر شبکهٔ عصبی دوسویه، ۸/۵ درصد بهبود برای همان نسبت سیگنال به نوفه به‌دست آمده است.

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

۱- مقدمه

با رشد روزافزون استفاده از سامانه‌های بازشناسی گفتار در کاربردهای عملی و روزمره، نیاز به حفظ کارایی بازشناسی گفتار در محیط‌های واقعی به‌عنوان امری اجتناب‌ناپذیر مطرح است. بنابراین هنگامی‌که از سامانه بازشناسی گفتاری که در محیط آزمایش‌گاهی آموزش داده شده است، در محیط واقعی استفاده شود اغلب کارایی سیستم بازشناسی، به‌دلیل عدم انطباق[1] دادگان آموزشی و دادگان محیط واقعی، به مقدار زیادی کاهش می‌یابد. ازاین‌رو بحث مقاوم‌سازی[2] در برابر نوفه به‌عنوان یکی از ضرورت‌های زمینه‌های فعّال تحقیقاتی در زمینهٔ بازشناسی مقاوم گفتار در سال‌های اخیر مطرح شده است که در آن لازم است سامانهٔ بازشناسی گفتار، به‌گونه‌ای منعطف و تطبیق‌پذیر طراحی گردد که امکان شناسایی گفتار نوفه‌ای 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 به‌طور معمول دو رویکرد کلّی وجود دارد. در رویکرد اوّل، جبران اثر نوفه در دادگان آزمون، با اصلاح بُردارهای بازنمایی گفتار صورت می‌پذیرد که آنها را روش‌های مبتنی‌بر جبران‌سازی دادگان می‌نامند. در رویکرد دوم هدف، اصلاح پارامترهای مدل طبقه‌بندی‌کننده برای جبران اثر نوفه است که روش‌های جبران‌ساز طبقه‌بندی‌کننده، نامیده می‌شوند.

[1]- Mismatch
[2]- Robustness

در روش‌های جبران‌سازی دادگان از قبیل نرمالیزه‌کردن کپستروم وابسته به کد [1] (CDCN) (*Acero, 1993*)، سری‌های تیلوربرداری [2] (VTS) (*Moreno, 1996*)، تفریق طیفی [3](*Boll, 1979*) و فیلتر وینر(*Porter and Boll, 1984*)، اثر نوفه روی دادگان، با تخمین طیف نویز جبران می‌گردد و سایر روش‌ها از قبیل جبران‌سازی کپستروم مبتنی بر تابع گوسی چندمتغیره [4] (*Moreno, 1996*) و فیلتر بهینه احتمالی [5] (POF) (*Neumeyer and Weintraub, 1994*)به تطبیق بین دادگانی که به‌طور هم‌زمان در شرایط آموزش و آزمون ضبط شده‌اند، می‌پردازند.

روش‌های جبران‌سازی طبقه‌بندی‌کننده از قبیل ترکیب مدل موازی [6] (PMC) (*Gales and Young, 1993*)و ترکیب مدل [7](*Varga and Moore, 1990*)، توزیع کلاس‌های آوایی را برای رسیدن به نوفهٔ جمعی، اصلاح می‌کنند و سایر روش‌ها از قبیل رگرسیون خطّی با بیش‌ترین احتمال [8] (MLLR) (*Leggetter and Woodland, 1994*)، مشخصات توزیع‌ها را برای متناسب‌شدن با گفتار نوفه‌ای آزمون، تغییر می‌دهند. یکی از موانع تمامی این روش‌ها در این است که همگی آنها، نوفه مورد نظر را ایستان فرض می‌کنند. تمامی این روش‌ها در مواجهه با نوفه ایستان با توان پایین یا متوسط، موفقیت به خوبی از خود نشان می‌دهند (گفتار نوفه‌ای با $SNR = 10\ dB$ 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 به‌طور کامل بی‌نتیجه خواهند بود (*Raj et al., 1997*).

یکی از روش‌های مؤثّر و کارآمد که در حوزهٔ جبران‌سازی دادگان آغشته به نوفهٔ ایستان و غیرایستان مطرح شده است، روش ویژگی‌های مفقود است (*Hennansky et al., 1996; Moore, 1997*) . در این روش برای تخمین مؤلّفه‌های مفقود اسپکتروگرام، از همبستگی بین مؤلّفه‌های طیفی گفتار در طول گفتار، استفاده می‌شود. در روش‌های مبتنی‌بر ویژگی‌های مفقود، نه تنها می‌توان مقدار *SNR* را در باند های فرکانسی متمایز، متفاوت دانست، بلکه می‌توان تخریب سیگنال گفتار در طول زمان را نیز تحت تأثیر سطوح مختلف نوفه در نظر گرفت. در این روش، نواحی با *SNR* پایین، از اسپکتروگرام گفتار پاک‌شده و با مقادیر تخمین بهینه‌ای از مقادیر مورد قبولشان جایگزین می‌شوند.

مزیّت روش ویژگی‌های مفقود در این است که در این روش، هیچ فرضی دربارهٔ ایستان‌بودن نوفهٔ مخرب درنظر گرفته نمی‌شود؛ هم‌چنین این روش، نیازی به اطّلاعات کامل از ساختار طیفی نوفه نداشته و تنها به توصیفی از نواحی نمایش زمانی- فرکانسی، با عنوان معتبر و نامعتبر بسنده می‌کند(*Bourlard and Dupont, 1996; Cooke et al., 1994; Moore, 1997*). روش مبتنی بر ویژگی‌های مفقود، از دو بخش اصلی شناسایی مؤلّفه‌های مفقود و اصلاح آنها تشکیل شده است.

ایدهٔ کارآمد دیگری که در حوزهٔ اصلاح بُردارهای بازنمایی گفتار مطرح شده، استفاده از شبکهٔ عصبی دوسویه [9] (BNN) است که به‌صورت هم‌زمان بر روی دادگان تمیز و نوفه‌ای تلفنی، آموزش داده شده است. در این روش پس از انجام یک سری پردازش‌های غیرخطّی و دوطرفه (جلوسو و برگشتی) از دانش نهفته در مدل، ناشی از یاد گرفتن گفتار تلفنی و نوفه‌ای بهره گرفته و بردارهای بازنمایی گفتار، در جهت افزایش صحّت بازشناسی آواهای گفتار بهبود داده شده‌اند (*Lippmann and Carlson, 1997*).

نقطهٔ قوت روش اصلاح ویژگی مبتنی بر شبکهٔ عصبی دوسویه نسبت به روش ویژگی‌های مفقود، این است که نیازی به شناسایی مؤلّفه‌های مفقود ندارد و بازسازی را در جهت هرچه شبیه‌تر شدن تمامی مؤلّفه‌های معتبر یا نامعتبر، به مؤلّفه‌های گفتار تمیز صورت می‌دهد و این یک برتری بسیار چشم‌گیری است که در این تحقیق حاصل شده است؛ چرا که درعمل، بحث شناسایی مؤلّفه‌های مفقود، که یک بحث چالش‌برانگیز را در تمامی روش‌های بازشناسی مقاوم گفتار مبتنی‌بر 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 آغشته به نوفه در سطوح مختلف پیاده‌سازی می‌شود و هم‌چنین الگوی ویژگی‌های مفقود با استفاده از شبکهٔ عصبی دوسویه در ادامهٔ تحقیقات قبلی ما به‌عنوان مسئلهٔ جبران‌سازی دادگان، مطرح خواهد شد که در آن عناصر مفقود در حوزهٔ طیف گفتار، به‌گونه‌ای بازسازی می‌شوند که حاوی اطّلاعات مفید در جهت بازشناسی آواهای گفتار باشند؛ بدون آنکه

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

9 -Bidirectional Neural Network (BNN)

نیاز به شناسایی دقیق مؤلّفه‌های مفقود طیف باشد. درنهایت دو روش با هم ترکیب‌شده و به‌صورت متوالی در اصلاح بردارهای بازنمایی گفتار نوفه‌ای ارزیابی خواهند شد. نتیجهٔ طراحی چنین تکنیک‌هایی مقاوم‌ترشدن سامانهٔ بازشناسی گفتار و حصول به نرخ‌های بالاتر بازشناسی در شرایط واقعی خواهد بود.

در ادامه در (بخش‌های دو و سه) مقاله، به‌ترتیب روش اصلاح بُردارهای بازنمایی گفتار مبتنی بر ویژگی‌های مفقود و شبکهٔ عصبی دوسویه مطرح خواهد شد. در بخش چهار دادگان مورد استفاده در این تحقیق و نحوهٔ استخراج ویژگی از آنها خواهد آمد. سپس در بخش پنج مقاله، آزمایش‌ها و نتایج حاصل از آنها ارائه خواهد شد و درنهایت در بخش شش، تحلیل نتایج صورت خواهد پذیرفت.

۲- روش ویژگی‌های مفقود در جبران‌سازی بردارهای بازنمایی نوفه‌ای

روش ویژگی‌های مفقود از دو بخش اصلی تشکیل می‌شود. بخش اوّل شناسایی مؤلّفه‌های مفقود طیف گفتار و بخش دوم بازسازی این مؤلّفه‌های مفقود از روی بخش‌های سالم گفتار است. در ادامه به تشریح کامل هر یک از این دو بخش خواهیم پرداخت.

۲-۱- بخش اوّل: شناسایی مؤلّفه‌های مفقود

یکی از مشکل‌ترین و حساس‌ترین بخش‌های سامانه‌های بازشناسی مقاوم گفتار مبتنی بر ویژگی‌های مفقود، تخمین ماسک‌های اسپکتروگرافیک است که مؤلّفه‌های نامعتبر طیف را شناسایی می‌کند *(Raj and Stern, 2005)*. این تخمین می‌تواند به چندین روش انجام شود، ممکن است SNR هر مؤلّفه طیف، به‌صورت مجزا برای شناسایی مؤلّفه‌های مفقود تخمین زده شود و یا اینکه مؤلّفه‌های مفقود، به‌طور مستقیم با استفاده از معیار دیگری غیر از SNR، تفکیک شوند. در این تحقیق با تخمین SNR هر مؤلّفهٔ طیف با استفاده از روش سری تیلور برداری، مؤلّفه‌های نامعتبر شناسایی خواهند شد.

تکنیک سری تیلور برداری[1](VTS) از جمله کاراترین روش‌های بازشناسی مقاوم گفتار است که در هر دو حوزهٔ اصلاح بردارهای بازنمایی و اصلاح مدل صوتی بازشناسی کاربرد دارد *(Moreno, 1996)*. در این تحقیق به شرح این تکنیک تنها در تخمین نوفهٔ بردارهای بازنمایی تخریب شده توسط نوفهٔ جمعی از روی سیگنال گفتار فارسی خواهیم پرداخت. بدین ترتیب که ابتدا بردارهای بازنمایی لگاریتم انرژی بانک فیلتر[2](LFBE) از گفتار نوفه‌ای استخراج شده و سپس به کمک تکنیک VTS میانگین نوفه جمعی برای تک‌تک باندهای فرکانسی به‌دست می‌آید.

سری تیلور برداری، یک الگوریتم جبران‌ساز نوفه است که به دنبال تخمین حداکثر شباهت از پارامترهای فیلتر خطی و نوفهٔ جمعی بر روی بردارهای لگاریتم طیفی سیگنال‌های نوفه‌ای است *(Moreno, 1996)*. اگر $Y(m)$ بیانگر بردار لگاریتم طیفی فریم mام یک گویش باشد که توسط فیلترینگ خطی و نوفهٔ جمعی تخریب شده باشد و همچنین $X(m)$ بیان‌گر سیگنال تمیز و عاری از نوفهٔ همان گویش باشد؛ در این صورت می‌توان رابطهٔ (۱) را بین این دو تعریف کرد *(Moreno, 1996)*:

$$Y(m) = X(m) + H + \log(1 + \exp(N - H - X(m))) \quad (۱)$$

که در آن H، لگاریتم مربع دامنهٔ طیف پاسخ ضربهٔ فیلتر خطی و N، لگاریتم طیف نوفه محسوب می‌شود. فرض بر این است که نوفه، ایستان بوده و تفاوت در بردارهای طیفی مجزای تخریب شده نوفه‌ای، مربوط به تفاوت‌های تحقق یک فرایند گوسی باشد. همچنین فرض بر این است که توزیع لگاریتم طیف نوفه در بردارهای مختلف، گوسی بوده و می‌توان نوفه را تنها با دو مؤلّفهٔ میانگین $\propto_N$ و واریانس Σ_N نمایش داد.

توزیع لگاریتم طیف سیگنال تمیز هم به‌عنوان یک فرایند گوسی مختلط فرض می‌شود. مجموعه پارامترهای مختلط گوسی Φ نیز از مجموعه دادگان گفتار تمیز به‌دست می‌آید

مسئله موجود در VTS در چارچوب رابطهٔ (۱)، برآورد احتمال حداکثر از پارامترکانال H، میانگین $\propto_N$ و واریانس Σ_N از طیف نوفه می‌باشد. درحقیقت اگر مجموعه‌ای از بردارهای طیفی سیگنال‌های نوفه‌ای را با Y نمایش دهیم، در روش VTS ما به دنبال حداکثر شباهت بین پارامترهای نوفهٔ سیگنال‌های نوفه‌ای و پارامترکانال H، میانگین $\propto_N$ و واریانس Σ_N از طیف نوفه هستیم. این برآورد را می‌توان به‌صورت رابطهٔ (۲) به نمایش گذاشت:

[1] - Vector Taylor Series (VTS)

[2] - Logarithm of Filter-Bank Energies (LFBE)

$$H, \propto_N, \Sigma_N = \arg\max_{H,\propto,\Sigma}\{P(Y|H, \propto, \Sigma, \Phi)\} \quad (۲)$$

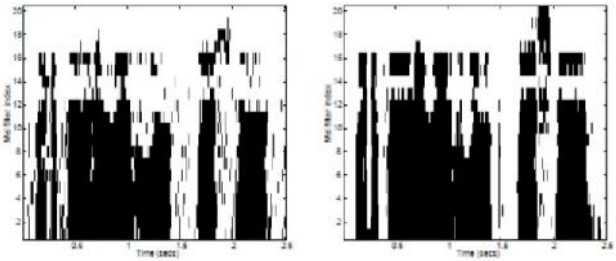
هنگامی که H، $\propto_N$ و Σ_N تخمین زده شد، $X(m)$ را می‌توان از $Y(m)$ با استفاده از تکنیک کمینه‌سازی میانگین مربعات خطا MMSE تخمین زد.

مقدار میانگین لگاریتم طیفی $\propto_N$ را می‌توان به‌عنوان مقدار واقعی لگاریتم طیفی نوفه نیز درنظرگرفت و از آن برای تخمین SNR محلی منحنی اسپکتروگرام سیگنال تخریب شده استفاده کرد.

اگر $Y(m,k)$ بیان‌گر مقدار مؤلّفهٔ kام باند فرکانسی از بردار طیفی mام اسپکتروگرام نوفه‌ای باشد و همچنین اگر kاُمین مؤلّفهٔ فرکانسی $\propto_N$ با $\propto_N(k)$ نمایش داده شود، مقدار تخمینی SNR از رابطهٔ (۳) به‌دست خواهد آمد:

$$\overline{SNR}(m,k) = \frac{Y(m,k) - \propto_N(k)}{\propto_N(k)} \quad (۳)$$

همان‌طور که در بخش قبل شرح داده شد، ماسک اسپکتروگرافیک بر مبنای SNR تخمینی محاسبه می‌شود؛ بدین صورت که تمامی مؤلّفه‌های اسپکتروگرام که مقدار $\overline{SNR}(m,k)$ آنها، از مقدار آستانهٔ T کم‌تر باشد، به‌عنوان



(شکل ۱): (سمت چپ) ماسک اسپکتروگرافیک به‌دست آمده از VTS برای یک گفتار نوفه‌ای تخریب شده با $SNR = 10\ dB$ (سمت راست) ماسک اسپکتروگرافیک از پیش دانسته برای همان گویش

مفقود برچسب‌دهی می‌شوند.

(شکل ۱) ماسک اسپکتروگرافیک تخمین‌زده شده با استفاده از روش VTS در کنار ماسک اسپکتروگرافیک ازپیش‌دانسته را که از روی اطّلاعات دقیق سیگنال نوفه افزوده شده به گفتار به‌دست آمده است را نشان می‌دهد.

۲-۲- روش 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 استخراج‌شده از گفتار، نمونه‌های یک فرایند تصادفی گوسی ایستان فرض می‌شوند. دانش اوّلیه دربارهٔ سیگنال‌های گفتار تمیز، به‌صورت پارامترهای آماری (مقدار میانگین بردارها و کواریانس بین مؤلّفه‌های بردارهای بازنمایی طیفی) نمایش داده می‌شوند.

μ_k نشان‌دهندهٔ میانگین kاُمین باند فرکانسی از mاُمین فریم بُردار طیفی $X(m,k)$، است. کواریانس بین k_1اُمین باند فرکانسی از mامین بردار طیفی $X(m,k_1)$ و k_2امین باند فرکانسی از $\xi + m$اُمین بُردار طیفی $X(m+\xi,k_2)$، نیز با $c(\xi,k_1,k_2)$ نمایش داده می‌شود که مقدار نرمالیزه شدهٔ نظیر آن $r(\xi,k_1,k_2)$ است که از طریق روابط زیر به‌دست می‌آیند:

(۴)

$$\propto(k) = E[X(m.k)]$$

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

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

در رابطهٔ فوق E[.] عمل‌گر امید ریاضی را نشان می‌دهد. این پارامترها از بردارهای طیفی استخراج شده از مجموعه دادگان تمیز به‌دست می‌آیند.

برای تخمین مؤلّفه‌های طیفی نامعتبر در m اُمین فریم بردار طیفی $X(m)$، همهٔ آنها را در بردار $X_u(m)$ گرد می‌آوریم؛ سپس همهٔ مؤلّفه‌های طیفی معتبر در نمایش اسپکتروگرام را که دارای کواریانس نرمالیزه شدهٔ حداقل ۰/۵ با حداقل یکی از مؤلّفه‌های بردار $X_u(m)$ باشند، شناسایی و در بردار $X_r^u(m)$ مرتب می‌کنیم. مقدار میانگین و کواریانس بین مؤلّفه‌های $X_u(m)$ و $X_r^u(m)$ و همچنین کواریانس متقابل بین آنها به‌وسیلهٔ پارامترهای میانگین و کواریانس حاصل از دانش اوّلیهٔ فرایند گوسی مورد نظر ساخته می‌شوند.

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

$$\hat{X}_u(m) = \propto_u + C_{ru}.C_{rr}^{-1}.(Y_r(m) - \propto_r) \quad (۵)$$

که در آن μ_u و μ_r به‌ترتیب نشان‌دهندهٔ مقدار میانگین مؤلّفه‌های $X_u(m)$ و $X_r^u(m)$ و همچنین C_{rr} و C_{ru} به‌ترتیب کواریانس و کواریانس متقابل بین مؤلّفه‌های $X_r^u(m)$ و $X_u(m)$ هستند. $\hat{X}_u(m)$ نیز معرف مقادیر بازسازی‌شدهٔ بردار ویژگی‌های مفقود نظیر $X_u(m)$ می‌باشد.

برای بازشناسی بهتر لازم است مقادیر بازسازی شدهٔ ویژگی‌های مفقود در یک حیطهٔ خاص، محدود شوند؛ که آن را در ترکیب با رابطه فوق دراصطلاح *Bounded MAP*

می‌نامند(*Raj, 2000*). رابطهٔ (۶) نحوهٔ اعمال این محدودیت را نشان می‌دهد.

$$\hat{X}_u(m,k)=\begin{cases}\hat{X}_u(m,k) & if\ \hat{X}_u(m,k)\le X_u(m,k)\\ X_u(m,k) & otherwise\end{cases} \quad (۶)$$

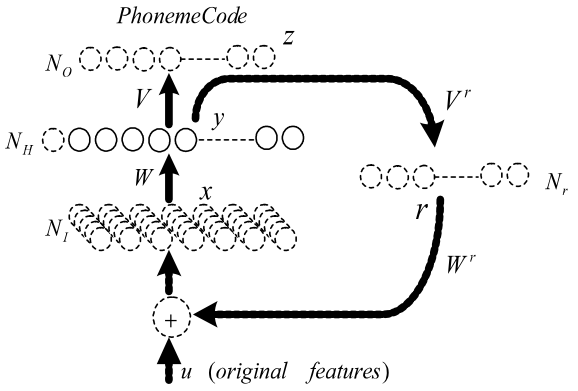
از آنجایی که سیگنال نوفه‌ای در اثر اضافه‌شدن نوفهٔ مخرب به گفتار تمیز، حاصل شده است، درنتیجه در رابطهٔ (۶) مقادیر لگاریتم طیفی سیگنال گفتار نوفه‌ای به‌عنوان یک کران بالا برای مقادیر بازسازی شده در نظر گرفته شده است.

۳- شبکهٔ عصبی دوسویه و اصلاح بردارهای بازنمایی گفتار

شبکهٔ عصبی دوسویه برای اوّلین‌بار 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 نوفه‌ای با نسبت‌های مختلف SNR استفاده می‌شود و با روش‌های رایج دادگان مفقود مقایسه خواهد شد. لذا در ابتدا ساختار و نحوهٔ عملکرد این نوع شبکه ارائه می‌شود.

۳-۱- ساختار شبکهٔ عصبی دوسویه

ساختار شبکهٔ عصبی دوسویه مطابق (شکل ۲) مشتمل بر یک شبکهٔ جلوسوی MLP با توابع غیرخطی تانژانت هایپربولیک در لایهٔ مخفی، به همراه یک شاخهٔ بازگشتی از لایهٔ مخفی آن به ورودی است. این شاخهٔ بازگشتی خود شامل یک لایهٔ مخفی با N_r نورون، تابع غیرخطی نوع تانژانت هایپربولیک و وزن‌های تمام متصل W^r و V^r است.



(شکل ۲): ساختار شبکهٔ عصبی دو سویه

خروجی شاخهٔ بازگشتی، خود یک بردار N_I مؤلّفه‌ای است که با مؤلّفه‌های نظیر خود از بردارهای بازنمایی ورودی u به‌صورت رابطهٔ (۷) ترکیب می‌شوند.

$$x_i[n]=\lambda u_i+\sum_{l=0}^{N_r} r_l[n]\ w^r_{li} \qquad i=1,2,...,N_I \quad (۷)$$

رابطهٔ (۷) نشان می‌دهد که حاصل جمع کسری $0<\lambda<1$ از بردار بازنمایی u، و یک ترکیب خطی از مقادیر لایهٔ مخفی بازگشتی شبکهٔ r، ورودی شبکه به‌ازای این بردار بازنمایی را در دورهٔ nاُم 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 لایهٔ مخفی بخش جلوسو، به‌ازای 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 شبکهٔ عصبی دوسویه یعنی مقادیر u_i ها بین [۱، ۱-] درکنار خروجی‌های لایهٔ مخفی بخش بازگشتی شبکه که در محدوده [۱، ۱-] هستند، برای هم‌گراشدن رابطهٔ (۷) ضروری است.

مقادیر لایهٔ مخفی بازگشتی شبکه، مطابق رابطهٔ (۸) خود به‌صورت یک تابع غیرخطی از روی مقادیر لایهٔ مخفی بخش جلوسوی شبکهٔ (y)، حاصل از دورهٔ $n-1$ تعلیم به‌دست می‌آیند. بدین ترتیب لازم است که برای کلیهٔ دادگان تعلیم مقادیر y نظیر آنها از دورهٔ فعلی به‌منظور استفاده در دورهٔ بعدی تعلیم شبکه نگهداری شوند.

$$r_l[n]=f\left(\sum_{j=0}^{N_H} y_j[n-1]v^r_{jl}\right) \qquad l=1,2,...,N_r \quad (۸)$$

$$y_j[n]=f\left(\sum_{i=0}^{N_I} x_i[n]w_{ij}\right) \qquad j=1,2,...,N_H \quad (۹)$$

$$z_k[n]=f\left(\sum_{j=0}^{N_H} y_j[n]v_{jk}\right) \qquad k=1,2,...,N_O \quad (۱۰)$$

هدف از به‌کارگیری مسیر بازگشتی در شبکهٔ BNN این است که از روی دانش یادگیری شده در لایه مخفی بخش جلوسوی این شبکه، مؤلّفه‌های بازنمایی تخریب شده یا مفقود شده در ورودی اصلاح شوند. دلیل فراهم شدن چنین امکانی در این شبکه مربوط به این است که شبکهٔ جلوسو برای هر آوای گفتار، بردارهای بازنمایی سالم از گفتار تمیز را در اختیار دارد که به علّت همبستگی بهتر و تعداد بیشتر آنها نسبت به بردارهای بازنمایی تخریب شده و پراکنده نویزی، آنها را بهتر یاد می‌گیرد. علاوه‌بر آن، دانش یادگرفته‌شده در لایهٔ مخفی شبکه، بیشتر حاوی اطّلاعات آوایی گفتار است و بسیاری از پراکندگی‌های بین ویژگی‌های ورودی که برای بازشناسی آواها مفید نمی‌باشند نیز در آن

سطح از بین رفته است. بنابراین برای بهبود و اصلاح بردارهای بازنمایی ورودی بهترین اطّلاعات در این سطح شبکه قرار گرفته است *(Vali et al., 2006a)*. در مسیر بازگشتی شبکهٔ BNN از روی این دانش، مؤلّفه‌های اصلاحی برای جمع شدن با هر یک از مؤلّفه‌های بردارهای بازنمایی ورودی توسط یک تابع غیرخطی تخمین زده می‌شوند. لازم به ذکر است که اگرچه این مؤلّفه‌های اصلاحی با مؤلّفه‌های بردارهای بازنمایی ورودی جمع می‌شوند، امّا به‌دلیل 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 دورهٔ جبران‌ساز ناشی از تخمین مسیر بازگشتی شبکه است.

تعلیم بردارهای بازنمایی نوفه‌ای و تمیز به‌صورت 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. بدین‌ترتیب برای هر دور تعلیم ضمن ساخته‌شدن اطّلاعات مفید بازشناسی در لایهٔ مخفی بخشِ 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 به‌صورت هم‌زمان تعلیم داده می‌شود. در این مرحله وزن‌های بازگشتی شبکهٔ 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 شبکهٔ مذکور به یک مقدار به‌طور تقریبی ثابت در رابطهٔ (۷) هم‌گرا شده و در عین حال صحّت بازشناسی در طی این دور زدن به‌صورت صعودی افزایش پیدا کرده تا به یک مقدار به‌طور تقریبی ثابت برسد.

۳-۲- اصلاح بردارهای بازنمایی توسط شبکه دوسویه

همان‌طور که اشاره شد شبکهٔ عصبی دوسویه (شکل ۲) از دو بخش اصلی تشکیل یافته است: یکی بخش جلوسو که شبیه به یک شبکهٔ عصبی MLP معمولی است و کار دسته‌بندی الگوها را انجام می‌دهد و دیگری بخش بازگشتی شبکه است که از لایهٔ مخفی بخش جلوسوی شبکه به ورودی شبکه بر می‌گردد و یک نگاشت از روی مقادیر لایهٔ 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 جمع‌شده و ورودی اصلاح‌شده 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 خوشه‌بندی مناسب و بازشناسی بهتر صورت گیرد. انتظار است که تعلیم بازنمایی‌های اصلاح‌شده به یک مدل بازشناسی جدید، منجر به بازشناسی بالاتری نسبت به مدل تعلیم‌داده‌شده بر روی دادگان اصلی شود.

مزیت این روش نسبت به روش‌هایی مشابه که تنها با تخمین خطّیِ اعوجاج‌ها تلاش برای بهبود بازنمایی‌های گفتار می‌کنند در این است که اوّل‌این‌که 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. دوم‌این‌که روش شبکه دوسویه یک روش اصلاحی وابسته به آواست، یعنی به عبارت دیگر، بردارهای بازنمایی گفتار در جهت بهبود صحّت بازشناسی آوا اصلاح شده‌اند *(Vali et al., 2006b)*.

۴- طراحی سامانهٔ بازشناسی گفتار

۴-۱- دادگان گفتار

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

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 به‌طور تقریبی تمامی واج‌های زبان فارسی را در بر می‌گیرند *(Bijankhan et al., 1994)*. بنابراین یک سیستم بازشناسی گفتار پیوسته مستقل از گوینده ولی وابسته به متن خواهد بود. علّت 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 تعمیم نتایج این تحقیق لطمه‌ای وارد نمی‌کند. بدین‌ترتیب 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 به‌طور کامل متفاوت هستند. دادگان نوفه‌ای مورد استفاده در این تحقیق، حاصل اضافه‌کردن نوفهٔ جمعی سفید گوسی[1] با نسبت سیگنال به نوفه‌های مختلف *(0dB, 5dB,10dB,15dB,20dB)* بر روی مجموعه دادگان تمیز فارس‌دات، به‌صورت دستی است. نرخ نمونه‌برداری سیگنال‌ها برابر ۱۶ کیلوهرتز می‌باشد.

۴-۲- استخراج بازنمایی LFBE و MFCC

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

پارامترهای لگاریتم انرژی بانک فیلتر[2] (LFBE) جزو چند روش متداولی است که از دستگاه شنوایی انسان الهام گرفته شده است. در این روش بازنمایی، فواصل فرکانسی مابین فیلترها و پهنای باند آن‌ها در مقیاس مل تنظیم می‌شوند که به‌طور تقریبی تا فرکانس یک کیلوهرتز به‌صورت خطی و بالاتر از این فرکانس به‌صورت لگاریتمی می‌باشد. در این تحقیق نیز به‌منظور 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 بازشناسی مقاوم گفتار چنانچه برای جبران‌سازی اثر نوفهٔ جمعی یا کانال از روی بردارهای بازنمایی گفتار، لازم باشد اثر تخریب بر روی هر یک از پارامترها به‌صورت مجزا شناسایی و برطرف 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 به‌دلیل 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 به‌دست آوردن طیف سیگنال از تبدیل فوریه استفاده می‌شود و چون تبدیل فوریه برای سیگنال‌های ایستان تعریف شده و سیگنال گفتار یک سیگنال غیرایستان و شبه‌دوره‌ای است، لذا باید از تبدیل فوریهٔ زمان‌کوتاه STFT استفاده شود *(Moulines and Verhelst, 1995)*. به‌طورمعمول طول پنجره و فریم انتخابی جهت تبدیل فوریهٔ زمان‌کوتاه بین ۱۰ تا ۳۰ میلی‌ثانیه اختیار می‌شود تا اطمینان حاصل شود که در طول یک فریم سیگنال به‌طور تقریبی حالت ایستان دارد. با توجّه به اینکه فرکانس نمونه‌برداری سیگنال‌های مورد استفاده ۱۶۰۰۰ هرتز می‌باشد، طول هر پنجرهٔ زمانی و گام پیش‌روی، به‌ترتیب ۲۵۶ و ۱۲۸ نمونه انتخاب شده است که از نظر زمانی شانزده میلی‌ثانیه بوده و همپوشانی فریم‌ها پنجاه درصد است. پنجرهٔ زمانی انتخاب شده از نوع همینگ است.

از آنجایی‌که در انتخاب بهینهٔ بردارهای بازنمایی گفتار از مشتقات اوّل 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 مؤلّفه است.

[1]- White Gaussian Additive Noise
[2]- Logarithm of Filter Bank Energies (LFBE)

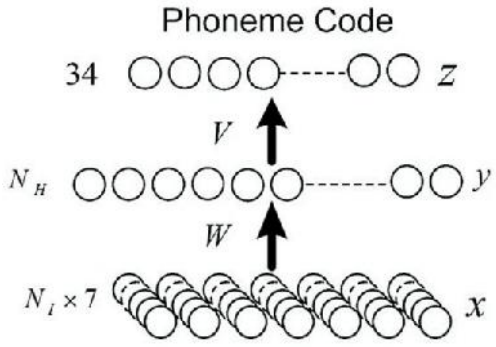
۴-۳- ساختار مدل بازشناسی

یکی از بخش‌های اصلی سامانه‌های بازشناسی گفتار پس از استخراج بردارهای بازنمایی مناسب از گفتار، مدل بازشناسی است. در یک تقسیم‌بندی کلی مدل بازشناسی می‌تواند آماری باشد که به‌طور عمومی مبتنی بر مدل مخفی مارکوف (HMM) است و یا یک مدل هوش مصنوعی باشد که به‌طور عمومی مبتنی بر شبکهٔ عصبیMLP هستند و یا یک مدل ترکیبی MLP/HMM باشد. از آنجایی‌که روش‌های ارائه‌شده برای بازشناسی 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 بازشناسی مرجع بر روی بردارهای بازنمایی گفتار تمیز تعلیم داده شود. سپس برای بررسی میزان کارایی روش‌های اصلاح بردارهای بازنمایی، مدل‌های بازشناسی جدید با ساختار شبیه به این مدل مرجع بر روی بردارهای بازنمایی اصلاح‌شدهٔ تعلیم‌داده، ارزیابی می‌شوند و نتایج آنها با مدل مرجع مقایسه خواهد گردید.

اساس مدل‌های بازشناسی گفتار به کار رفته در این تحقیق بر مبنای نوع خاصی از شبکهٔ عصبی MLP است که با لحاظ‌کردن چند فریم متوالی به‌عنوان ورودی، شبیه به ساختار شبکه‌های عصبی TDNN[1] می‌شود (*A. Waibel et al., 1989*). لذا در این بخش، ساختار بهینهٔ شبکهٔ عصبیMLP برای دادگان گفتاری این تحقیق، شناسایی و طراحی و براساس آن مدل بازشناسی بر روی بردارهای بازنمایی گفتار تمیز تعلیم داده می‌شود تا در ادامه به‌عنوان مرجع مقایسه از آن استفاده شود.

(شکل۳) ساختار نورونی یک شبکهٔ عصبی MLP را با یک لایهٔ پنهان به‌صورت $N_I \cdot 7 \rightarrow N_H \rightarrow 34$ 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 $N_I \cdot 7$ نورون خواهد شد. خروجی شبکه نیز به تعداد کدهای آوایی موجود در دادگان تعلیم فارسی یعنی ۳۴ نورون در نظر گرفته شده است.

شبکهٔ ذکرشده بر روی ویژگی‌های MFCC استخراج شده از ویژگی‌های LFBE ، تعلیم داده می‌شود که تعداد نورون‌های لایهٔ مخفی آن یکصد نورون انتخاب شده است. این تعداد نورون بر اساس حجم دادگان تعلیمی شبکه، به‌صورت بهینه انتخاب شده است(*Blumer et al., 1989*). مدل بازشناسی، بر روی بردارهای بازنمایی دادگان گفتاری تمیز تا همگرایی کامل تعلیم داده می‌شود. الگوریتم تعلیم پس‌انتشار خطا با معیار همگرایی کاهش شیب منحنی خطای تعلیم از یک مقدار آستانه است.



(شکل ۳): مدل بازشناسی مرجع مبتنی بر شبکهٔ عصبی TDNN

۵- شرح آزمایش‌ها

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

از آنجایی که در این تحقیق چندین‌بار بردارهای بازنمایی گفتار تمیز و نوفه‌ای با چندین روش مختلف اصلاح می‌شوند برای ارزیابی میزان موفقیت روش‌های پیشنهادی در اصلاح بردارهای بازنمایی، لازم است مدل‌های بازشناسی مجزایی 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 نوفه‌ای 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 نوفهٔ سفید گوسی جمعی تخریب شده است. (جدول۱) صحت بازشناسی مدل مرجع برای دادگان آزمون گفتار تمیز و نوفه‌ای قبل از هرگونه اصلاح دادگان را نشان می‌دهد.

(جدول ۱): صحّت بازشناسی برای دادگان آزمون گفتار تمیز و نویزی در مدل بازشناسی 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

[1] Time Delay Neural Network (TDNN)

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

در این بخش روش بازسازی مبتنی بر کواریانس برای جبران اثر نوفهٔ جمعی ارزیابی خواهد شد. این ارزیابی طی دو مرحله صورت خواهد گرفت. در مرحلهٔ اوّل صرفاً ارزیابی صحت بازسازی مؤلّفه‌ها مورد نظر است. امّا مرحلهٔ دوم ارزیابی بر روی صحت بازشناسی است که ابتدا شامل شناسایی مؤلّفه‌های مفقود و سپس بازسازی آنها خواهد بود. بردارهای بازنمایی گفتار، LFBE*های* ۲۰ مؤلّفه‌ای هستند که پس از بازسازی‌شدن مؤلّفه‌های مفقود آنها، با اعمال تبدیل DCT به مؤلّفه‌های MFCC تبدیل می‌شوند.

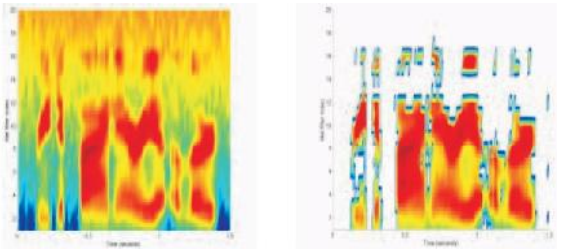
در مرحلهٔ اوّل برای ارزیابی صحّت بازسازی، نیازی به شناسایی مؤلّفه‌های مفقود نیست؛ چرا که مؤلّفه‌های مفقود با دانش قبلی از نوفه تعیین می‌شوند که به آن ارزیابی از روی ماسک ازپیش‌دانسته[1] می‌گوییم. در روش ماسک ازپیش‌دانسته، اسپکتروگرامی که مؤلّفه‌های مفقود آن از قبل شناسایی و پاک شده‌اند، برای ارزیابی صحبت بازسازی استفاده می‌شود و این تنها تفاوت بین روش ماسک ازپیش‌دانسته و روش ارزیابی کلّی است. این روش می‌تواند یک ارزیابی اولیه مفیدی از عملکرد روش بازسازی مبتنی بر کواریانس محسوب شود.

هنگامی‌که سیگنال گفتار توسط یک نوفهٔ جمعی غیر همبسته تخریب می‌شود، طیف توان سیگنال حاصل از جمع طیف توان گفتار تمیز با طیف توان نوفه طبق رابطهٔ (۱۱) به‌دست می‌آید. زمانی‌که طیف توان نوفهٔ جمعی N_p را در اختیار داشته باشیم، از این طریق می‌توان طیف توان سیگنال تمیز X_p را از طیف توان سیگنال نوفه‌ای Y_p به‌دست آوریم و سپس می‌توان SNR را به‌صورت محلی از رابطهٔ (۱۲) به‌دست آورد:

$$Y_p(m,k) = X_p(m,k) + N_p(m,k) \qquad (۱۱)$$

$$SNR(m,k) = 10\log 10\frac{X_p(m,k)}{N_p(m,k)} \qquad (۱۲)$$

که $X_p(m,k)$ و $N_p(m,k)$ به‌ترتیب طیف توان گفتار تمیز و نوفه را در باند فرکانسی kاُم از فریم mاُم اسپکتروگرام نشان می‌دهند. حال می‌توان تمامی مؤلّفه‌های اسپکتروگرام را که از یک حدّ آستانهٔ ثابت، SNR پایین‌تری را داشته باشند، به‌عنوان مؤلّفه‌های مفقود درنظرگرفت و از بردارهای بازنمایی $LFBE$ حذف کرد. چون مقدار ثابتی برای حدّ آستانه وجود ندارد، ما در این تحقیق، حدّ آستانه‌های متفاوتی را مورد ارزیابی قرار دادیم و بهترین حدّ آستانه را که در ارزیابی‌ها به‌خوبی مؤلّفه‌های مفقود را از مؤلّفه‌های معتبر تفکیک می‌کرد انتخاب کردیم. (شکل ۴) اسپکتروگرام یک گفتار نوفه‌ای و ماسک اسپکتروگرام همان گفتار نوفه‌ای را زمانی که تمامی مؤلّفه‌های با SNR پایین‌تر از حد آستانه $3dB$ پاک شده‌اند، نمایش می‌دهد.



(شکل ۳): سمت چپ، اسپکتروگرام یک گفتار نوفه‌ای و شکل سمت راست، ماسک اسپکتروگرام همان گفتار نوفه‌ای زمانی که تمامی مؤلّفه‌های با SNR پایین‌تر از حد آستانه 3dB از اسپکتروگرام پاک شده‌اند ***(RAJ, 2000)***.

در ارزیابی اوّلیه مشاهده شد که روش مبتنی بر کواریانس عملکرد چشم‌گیری در بازسازی نواحی مفقود داشته است. با این اوصاف انتظار می‌رفت که بهبود قابل ملاحظه‌ای نیز در بازشناسی گفتارهای تخریب‌شده توسط نویز با SNR*های* مختلف داشته باشد، لذا در (شکل۵) نتایج بازشناسی روی بردارهای بازنمایی MFCC اصلاح شده توسط روش مبتنی بر کوواریانس در روش ماسک ازپیش‌دانسته ارائه شده است. مدل بازشناسی، همان شبکه مرجع است که با دادگان تمیز تعلیم یافته و با دادگان اصلاح شده در SNRهای مختلف ارزیابی صورت گرفته است. نتایج به‌دست آمده، قابلیت روش را در بازسازی نواحی تخریب شده توسط نوفه، به‌خوبی نشان می‌دهد. این بهبود (در صحّت بازشناسی) حداقل ۲/۳ درصد برای بازسازی گفتارهای نوفه‌ای با SNR=20dB و حداکثر ۱۵/۲ درصد برای بازسازی گفتارهای تخریب‌شده توسط نوفه با SNR=0dB، نمود پیدا کرد.

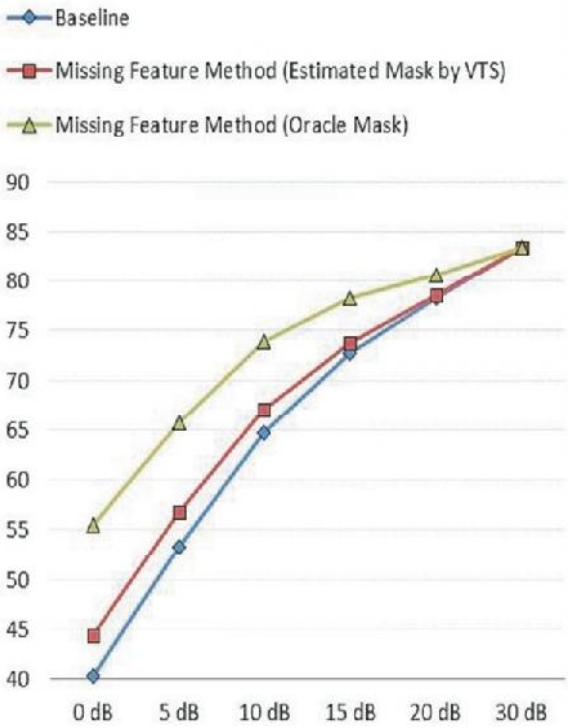
درعمل پیاده‌سازی چنین سامانه‌ای میسر نیست چرا که هیچ دانشی از میزان نوفهٔ تخریبی و نواحی تخریب شده توسط آن در دسترس نیست و باید سامانه، طوری طراحی شود که به‌صورت خودکار قادر به شناسایی مؤلّفه‌های مفقود اسپکتروگرام باشد.

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

[1] Oracle Mask

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

همان‌طورکه (شکل۵) نشان می‌دهد زمانی‌که نواحی تخریب‌شدهٔ اسپکتروگرام از قبل شناخته شده باشند، نتایج صحّت بازشناسی به‌دست آمده از روش‌های بازسازی، خیلی بالاتر از زمانی است که از روش مبتنی بر کواریانس در حوزهٔ ویژگی‌های مفقود استفاده شود؛ با این وجود، صحّت بازشناسی ۴۴/۵ درصدی و بهبود ۴/۲ درصدی بدست آمده برای بازسازی گفتارهای تخریب‌شده توسط نوفهٔ با SNR=0dB نشان می‌دهد که عملکرد روش کواریانس در شناسایی و بازسازی نواحی مفقود و بالا بردن صحت بازشناسی مؤثّر بوده و می‌توان از این روش به‌عنوان یکی از روش‌های موفق در جبران‌سازی اثر نوفه استفاده کرد.



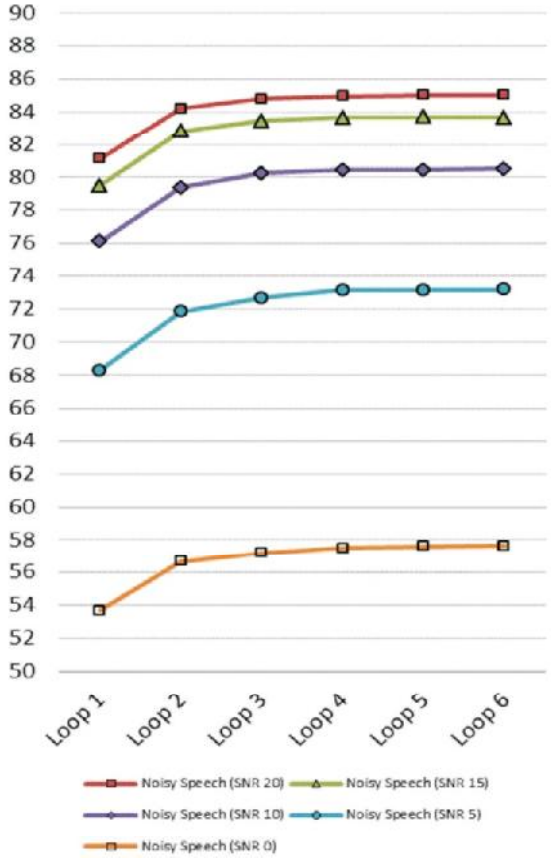
(شکل ۴): صحّت بازشناسی به‌دست آمده از اصلاح بردارهای بازنمایی MFCC با استفاده از روش دادگان مفقود با ماسک از پیش دانسته و تخمین ماسک

۵-۳- اصلاح بردارهای بازنمایی با شبکهٔ عصبی دوسویه

در این بخش، توانایی شبکهٔ عصبی دوسویه برای اصلاح غیرخطی بردارهای بازنمایی *LFBE* مورد بررسی قرار می‌گیرد. نحوهٔ تعلیم شبکهٔ دوسویه، طبق روابط آمده در (بخش۳-۱) است. یعنی وزن‌های بخش جلوسو و بازگشتی شبکه، به‌صورت هم‌زمان در هر دور تعلیم، اصلاح می‌شوند.

در مرحلهٔ آزمون شبکه، دادگان آزمون مطابق روابط (۷) تا (۱۰) برای $n=1,2,\ldots,6$ دور در شبکهٔ چرخانده‌شده و هر بار صحّت بازشناسی در خروجی شبکهٔ دوسویه برای آنها به‌دست آمده است. $n=1$ به منزله این است که فقط بخش جلوسوی شبکهٔ دوسویه در بازشناسی دخالت داده شده است. پس از اوّلین دور مقادیر خروجی نورون‌های لایهٔ مخفی بخش جلوسوی شبکه (لایهٔ y) برای تمام دادگان آزمون مشخص شده و از آنها برای اصلاح ورودی و بازشناسی در دور دوم استفاده شده است. نتایج صحت 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 نوفه‌ای حاصل می‌شود. از تکرار $n=4$ به بعد نتایج به‌طور تقریبی ثابت و پایدار شده‌اند. لذا شبکهٔ 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 به‌طور تقریبی متوقف شده است و بازنمایی‌های جدید تولید شده در رابطهٔ (۷) را بتوان بعداً در مدل‌های بازشناسی استفاده کرد. تا اینجا هر یک از شبکه‌های دوسویهٔ پنج‌گانهٔ بالا بر روی یک دسته دادگان با SNR مشخّص 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 با هر SNR دلخواه بتواند بدون اینکه نیاز باشد همه دادگان با انواع SNR به یک شبکهٔ عصبی دوسویه تعلیم داده شود را در یک فرآیند پیشنهادی اصلاح کند. برای این منظور پیشنهاد زیر مطرح می گردد:

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



(شکل ۵): میزان افزایش صحّت بازشناسی 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 شبکهٔ دوسویه برای استفاده جهت اصلاح کلیهٔ بردارهای بازنمایی با SNR های مختلف موفق‌تر عمل کرده است، پنج مدل بازشناسی مبتنی بر شبکهٔ عصبی MLP شبیه به مدل مرجع (درقبل بر روی بردارهای بازنمایی تمیز تعلیم داده شده بود) بر روی بازنمایی‌های MFCC تمیز اصلاحی استخراج شده از هر یک از پنج شبکهٔ دوسویه، تعلیم داده می‌شود. (جدول ۲) نتایج بازشناسی گفتار بر روی دادگان آزمون اصلاحی با انواع SNR را در پنج مدل بازشناسی مذکور نشان می‌دهد.

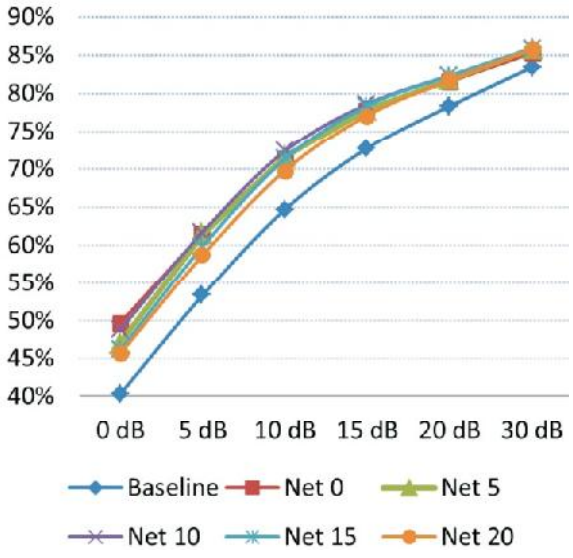
(جدول ۲): صحّت بازشناسی به‌دست آمده از اصلاح بردارهای بازنمایی به روش شبکه عصبی دوسویه برای بردارهای بازنمایی MFCC

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 شبکهٔ عصبی دوسویه پیشرفت قابل قبولی در بهبود بازنمایی‌های نوفه‌ای در تمام *SNR* ها داشته‌اند که البته این بهبود در *Net 5*، یعنی زمانی که از گفتار تمیز و نوفه‌ای با نسبت سیگنال به نوفهٔ *5dB* به‌طور هم‌زمان در تعلیم شبکهٔ عصبی دوسویه استفاده شده، بیشتر نمود پیدا کرده است که منجر به افزایش ۱/۷ درصدی در بازشناسی سیگنال تمیز و همچنین۲/۶ الی هشت درصدی در بازشناسی گفتارهای تخریب شده توسط نوفه با *SNR=20dB* الی *SNR=0dB* شده است. با وجود بهبود قابل ملاحظه در *Net 5*، (جدول ۲) به‌طور واضح نشان می‌دهد که در برخی از شبکه‌های دوسویه، در برخی از *SNR*ها، بهبودی در صحت بازشناسی حاصل نشده است و یا این بهبود چندان رضایت بخش نیست.

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

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



(شکل ۶): صحّت بازشناسی به‌دست آمده از اصلاح به روش شبکهٔ عصبی دوسویه برای بردارهای بازنمایی MFCC استخراج شده از بردارهای بازنمایی شصت‌تایی LFBE

پس از تعلیم پنج شبکهٔ دوسویه (Net x, x=0,5,10,15,20) و اصلاح کلیهٔ بردارهای بازنمایی شصت مؤلّفه‌ای LFBE در کلیه SNR ها و سپس تبدیل بردارهای LFBE به MFCC و تعلیم مجدد پنج مدل بازشناسی مبتنی بر ساختار مرجع بر روی کلیهٔ دادگان اصلاح‌شده (کلیهٔ SNR ها) در هر یک از Net x ها، نتایج صحّت بازشناسی برای دادگان آزمون اصلاحی در SNRهای مختلف به‌دست آمد که در نمودار (شکل ۷) گزارش شده است. همان‌طورکه مشاهده می‌شود در اصلاح بردارهای بازنمایی، استفاده از مقادیر دینامیک پارامترها در کنار ایستای آنها نقش مؤثّری در ارتقای صحّت بازشناسی داشته است و هر پنج شبکه به‌صورت محسوسی نتایج را برای کلیهٔ SNRها ارتقا داده‌اند چراکه هر چه شبکهٔ دوسویه در شناسایی آواها عملکرد بهتری داشته باشد دانش آوایی نهفته در لایهٔ 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 شبکهٔ دوسویه فوق در اصلاح برداهای بازنمایی مربوط به SNR نظیر خود موفق‌تر از سایر شبکه‌ها عمل کرده است زیرا طبیعی است که دادگان تعلیمی به یک شبکهٔ دوسویه بهتر از سایر دادگانی که در زمان تعلیم مشاهده نشده‌اند اصلاح خواهند شد.

بهترین نتایج مربوط به *Net10* است که بهبود ۴/۰ و ۸/۵ درصدی در صحّت بازشناسی گفتارهای نوفه‌ای به‌ترتیب برای *SNR=20dB* و *SNR=0dB* داشته است. بدین‌ترتیب نتایج نمودارهای (شکل ۷) هم عملکرد چشم‌گیر شبکه‌های عصبی دوسویه در بازسازی مؤلّفه‌های مفقود را نشان می‌دهد و هم بیان می‌کند زمانی که شبکه به‌وسیلهٔ بردارهای بازنمایی شصت‌تایی *LFBE* تعلیم داده شود میزان بهبود بیشتر و قابل اطمینان تر خواهد بود.

برای مقایسهٔ بهتر، دو روش پیشنهادی اصلاح ویژگی‌های ایستا و پویا با استفاده از شبکهٔ دوسویه، در (جدول ۳) میزان میانگین بهبود در صحّت بازشناسی برای کلیهٔ *SNR*ها آورده شده است.

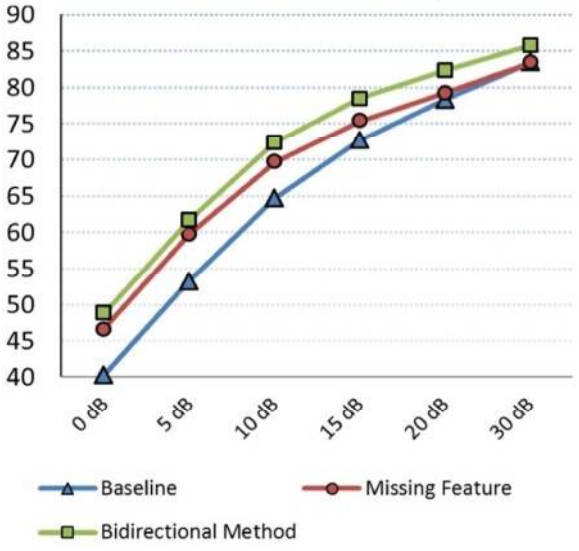
(جدول ۳): میزان میانگین بهبود صحّت بازشناسی در دو روش پیشنهادی استفاده از شبکهٔ عصبی دوسویه

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 توجّه به (جدول ۳) می‌توان ادعا کرد زمانی‌که از پویای بردارهای بازنمایی در کنار ایستای آنها (بردارهای بازنمایی شصت‌تایی LFBE) برای تعلیم شبکهٔ عصبی دوسویه استفاده شود، به‌کارگیری گفتارهای نوفه‌ای با سطح نوفه نزدیک به 10 dB در کنار گفتار تمیز برای تعلیم به شبکهٔ دوسویه مناسب است.

بدین‌ترتیب نحوهٔ استفادهٔ بهینه از شبکهٔ دوسویه برای اصلاح بردارهای بازنمایی، به‌صورت زیر خلاصه می‌شود:

- ابتدا از کلیهٔ بردارهای بازنمایی گفتار تمیز و نوفه‌ای در سطوح مختلف SNR، بردارهای بازنمایی بیست مؤلّفه‌ای LFBE استخراج می‌شوند.
- با اضافه‌کردن مشتقات اوّل و دوم بازنمایی در کنار مؤلّفه‌های ایستا، بازنمایی‌های شصت مؤلّفه‌ای ساخته می‌شوند.
- یک شبکهٔ عصبی دوسویه بر روی دادگان تمیز و نوفه‌ای با SNR در محدودهٔ صفر تا ده دسی بل تعلیم داده می شود.
- کلیهٔ دادگان تمیز و نوفه‌ای در سطوح مختلف SNR در طی چهار دور گردش در شبکهٔ دوسویهٔ ذکرشده اصلاح می‌شوند.
- دادگان تمیز اصلاح‌شده از بالا برای تعلیم به هر مدل بازشناسی گفتار قابل استفاده‌اند که بر روی تمام دادگان نوفه‌ای اصلاح شده ذکرشده نیز قابل ارزیابی خواهد بود.



(شکل ۷): صحّت بازشناسی به‌دست آمده از روش‌های دادگان مفقود و شبکهٔ عصبی دوسویه

۶- جمع بندی

هدف از این تحقیق، ارائهٔ یک راهکار جدید در بازشناسی مقاوم گفتار بوده است که تا جای ممکن روش ارائه شده برای جبران‌سازی بیشتر تنوعات گفتار، جواب‌گو باشد. شیوهٔ کلّی مواجه با بازشناسی مقاوم گفتار در بهبود و اصلاح بازنمایی‌های گفتار قبل از ورود به مدل بازشناسی بوده است. زیرا همان‌طور که می‌دانیم بهترین روش‌های دسته‌بندی الگو نیز در مقابل الگوهای تخریب‌شده که اطّلاعات مفید و

متمـایز کننـده را از دسـت داده باشـند، توانـایی مطلـوبی نخواهند داشت. نقطهٔ قوّت روش اصـلاح ویژگـی مبتنـی بـر شبکهٔ عصبی دوسویه نسبت به روش ویژگی‌های مفقود، این است کـه نیـازی بـه شناسـایی مؤلّفـه‌هـای مفقـود نـدارد و بازسازی را در جهت هرچه شبیه‌ترشدن تمـامی مؤلّفـه‌هـای معتبر یا نامعتبر به مؤلّفه‌های گفتار تمیز صورت می‌دهـد و این یک برتری بسیار چشم‌گیری است کـه در ایـن تحقیـق حاصل شده است؛ چرا که درعمل، بحث شناسایی مؤلّفه‌های مفقود، که یـک بحـث چـالش‌برانگیـز در تمـامی روش‌هـای بازشناسی مقاوم گفتار مبتنی بر ویژگی‌های مفقـود اسـت و ارتباط مستقیمی با میزان صـحت بازشناسـی دارد را حـذف می‌کند. ایدهٔ اصلی نهفته در تمام روش‌های ارائه شده در این تحقیق در استفاده از دانش مشترکی است کـه در بردارهـای بازنمایی گفتار تمیز و نوفه‌ای وجود دارد و می‌توان اطّلاعات مفیدی را که در اثر نوفه از بین رفته اسـت از بازنمـایی‌هـای سالم بیرون کشید و در بهبود بازنمایی‌های نوفـه‌ای بـه کـار گرفت.

بررسی و مقایسهٔ کارایی دو نوع روش جبران اثر نوفه، دادگان مفقود و شبکه‌های عصبی دوسویه در مدل بازشناسی بر روی گفتـار نوفـه‌ای و تمیـز و همچنـین کـاربرد تکنیـک شبکه‌های عصبی دوسویه در اصـلاح بردارهـای بازنمـایی در جهت نزدیک‌کردن بازنمایی‌های دسته‌های آوایـی یکسـان از دو نـوع گفتـار نوفـه‌ای و تمیـز بـه هـم‌دیگـر و دورکـردن بازنمـایی‌هـای دسـته‌هـای آوایـی متفـاوت از هـم‌دیگـر از نوآوری‌های این تحقیق به‌شمار می‌رفـت کـه شـرح آن داده شد.

پیشرفت چشم‌گیری که در روند بازسازی با استفاده از روش جبران‌سـازی توسـط شـبکه‌هـای عصـبی دوسـویه در مقایسـه بـا بازسـازی مؤلّفـه‌هـای مفقـود مبتنـی بـر روش کواریانس حاصل شد، در نمودار (شکل ۸) بـه وضـوح دیـده می‌شود.

(شکل ۸): به خوبی نشـان مـی‌دهـد کـه صـحّت بازشناسـی به‌دست آمده از روش جبران‌سازی توسط شبکه‌هـای عصـبی دوسویه در بهترین و بدترین حالت به‌ترتیب بهبـود ۸/۵ و ۴ درصدی نسبت به حالت پایه داشته است که معرف سـیگنال نوفه‌ای بدون هیچ جبران‌سازی اثر نوفه مـی‌باشـد. همچنـین بهبود ۴/۳ و ۳/۷ درصدی نیز نسبت به روش جبران اثر نوفهٔ مبتنی بر کواریـانس در SNRهـای مختلـف بـه‌دسـت 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.

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: 10.1 10 9/T AS SP . 1979 . 11632 09.

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. 83 7-840.

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-INFENG/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.

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, Pittsbrgh, 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.

حجت 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 ۱۳۸۹در پژوهش‌گاه مخابرات و الکترونیک نصر در گروه رمز مشغول به فعّالیت است. حوزه‌های تحقیقاتی تخصصی وی عبارتند از: پردازش سیگنال، پردازش گفتار، پردازش دادهٔ رمز، بازشناسی الگو، شبکه‌های عصبی، ماشین بردار پشتیبان.

نشانی رایانامهٔ 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 دانشکدهٔ برق دانشگاه صنعتی خواجه نصیرالدین طوسی می باشد. زمینه‌های پژوهشی مورد علاقهٔ ایشان پردازش گفتار، پردازش صوت و کاربرد آن در مهندسی پزشکی و شبکه‌های عصبی مصنوعی و زیستی است.

نشانی رایانامهٔ ایشان عبارت است از

mansour.vali@eetd.kntu.ac.ir