ارائهٔ یک روش جدید بهسازی گفتار بر مبنای یادگیری مدل ناهمدوس به‌کمک ضرایب تبدیل موجک

سمیرا مودّتی

استادیار، دانشکده فنی مهندسی، دانشگاه مازندران

چکیده

بهسازی گفتار یکی از زمینه‌های پرکاربرد در پردازش سیگنال است که در حوزه‌های مختلفی مورد استفاده قرار می‌گیرد. در این مقاله از مفاهیم بازنمایی تُنُک و یادگیری واژه‌نامه به‌منظور حذف نوفه از سیگنال گفتار در فضای ویژگی تبدیل موجک استفاده می‌شود. ساختار مورد نیاز جهت بازنمایی هر مؤلفه از سیگنال به‌کمک مفاهیم بازنمایی تُنُک، براساس تعداد کمی از اتم‌های یادگیری‌شده امکان‌پذیر است. به‌منظور دست‌یابی به نتایج مطلوب در بهسازی گفتار، از روال یادگیری واژه‌نامه ناهمدوس بهره گرفته می‌شود. به‌کمک ضرایب تبدیل موجک، تجزیه سیگنال در زیرباندهای مختلف که شامل اطلاعات دقیقی از محتوای سیگنال هستند، فراهم می‌شود. در روش پیشنهادی، دو سناریوی نظارت‌شده و نیمه‌نظارت‌شده مورد بررسی قرار گرفته و یک الگوریتم آشکارساز فعالیت گفتاری در هر سناریو با توجه به شرط‌های معرفی‌شده بر اساس واژه‌نامه‌های یادگیری‌شده در گام آموزش، پیشنهاد می‌شود. با استفاده از نتایج خروجی آشکارساز پیشنهادی، سیگنال گفتار تخمینی طی یک روال بهسازی در گام بعد به‌دست خواهد آمد. نتایج گزارش‌شده براساس معیارهای مختلف ارزیابی عملکرد، بر توانایی این روش در زمینه کاهش نوفه سیگنال گفتار تأکید می‌کند. روش‌های پیشنهادی، توانایی بالایی را درخصوص کاهش نوفه‌های ناایستا به‌خصوص در مقادیر سیگنال به نوفه پایین دارد.

واژگان کلیدی: بهسازی گفتار، بازنمایی تُنُک، واژه‌نامه ناهمدوس، تبدیل موجک، آشکارساز فعالیت گفتار.

A New Method for Speech Enhancement Based on Incoherent Model Learning in Wavelet Transform Domain

Samira Mavaddati

Electrical Department, Faculty of Technology and Engineering,
University of Mazandaran, Babolsar, Iran

Abstract

Quality of speech signal significantly reduces in the presence of environmental noise signals and leads to the imperfect performance of hearing aid devices, automatic speech recognition systems, and mobile phones. In this paper, the single channel speech enhancement of the corrupted signals by the additive noise signals is considered. A dictionary-based algorithm is proposed to train the speech and noise models for each subband of wavelet decomposition level based on the coherence criterion. Using the presented learning method, the self-coherence measure between different atoms of each dictionary and mutual coherence between the atoms of speech and noise dictionaries are minimized and lower sparse reconstruction error is yielded. In order to reduce the computation time, a composite dictionary is utilized including only the speech dictionary and one of the noise dictionaries selected corresponding to the noise condition in the test environment. The speech enhancement algorithm is introduced in two scenarios, supervised and semi-supervised situations. In each

* Corresponding author

*نویسندهٔ عهده‌دار مکاتبات

• تاریخ ارسال مقاله: ۱۳۹۶/۱۱/۱۸ • تاریخ پذیرش: ۱۳۹۸/۰۳/۲۹ • تاریخ انتشار: ۱۳۹۹/۰۹/۱۵ • نوع مطالعه: پژوهشی

[DOI: 10.29252/jsdp.17.3.17]

scenario, a voice activity detector (VAD) scheme is employed based on the energy of sparse coefficient matrices when the observed data is coded over the related dictionary.
The presented VAD algorithms are based on the energy of the coefficient matrices in the sparse representation of the observation data over the specified dictionaries. These speech enhancement schemes are different in the mentioned scenarios. In the proposed supervised scenario, domain adaptation technique is employed to transform a learned noise dictionary into an adapted dictionary according to the noise conditions of the test environment. Using this step, the observed data is sparsely coded with low sparse approximation error based on the current situation of the noisy environment. This technique has a prominent role to obtain better enhancement results particularly when the noise signal has non-stationary characteristics. In the proposed semi-supervised scenario, adaptive thresholding of wavelet coefficients is carried out based on the variance of the estimated noise for each frame in different subbands. These implementations are carried out in two different conditions, the training and test steps, as speaker dependent and speaker independent scenarios.
Also, different measures are applied to evaluate the performance of the presented enhancement procedures. Moreover, a statistical test is used to have a more precise performance evaluation for different considered methods in the various noisy conditions. The experimental results using different measures show that the presented supervised enhancement scheme leads to much better results in comparison with the baseline enhancement methods, learning-based approaches, and earlier wavelet-based algorithms. These results have been obtained for an extensive range of noise types including the structured, unstructured, and periodic noise signals in different SNR values.

Keywords: Speech enhancement, Dictionary learning, Sparse representation, Domain adaptation, Voice activity detector, Wavelet transform

۱- مقدمه

بهسازی گفتار یکی از زمینه‌های مورد علاقهٔ پژوهش‌گران جهت معرفی راه‌کارهای جدید در میان انواع حوزه‌های پردازش سیگنال است. انواع نوفهٔ محیطی می‌توانند به‌صورت جدی کیفیت و قابلیت فهم سیگنال گفتار را با مشکل مواجه کنند. دسته روش‌های پایه به‌منظور بهسازی گفتار ارائه‌شده تاکنون شامل الگوریتم‌های مبتنی بر تفاضل طیفی، طراحی فیلتر، روش‌های آماری، الگوریتم‌های زیرفضا و غیره است [1-4]. در سال‌های اخیر مفهوم تُنُکی[1] نیز در زمینه‌های مختلف پردازش سیگنال مورد توجه قرار گرفته است. این مفهوم بیان‌گر نمایشی از سیگنال است که در آن قسمت عمده‌ای از اطلاعات اساسیِ سیگنال در تعداد کمی از ضرایب متمرکز شده است. تُنُکی در حالت ایده‌آل به‌معنای صفربودن بیش‌تر مؤلفه‌های سیگنال بوده، اما درعمل تُنُکی به‌معنای آن است که بیش‌تر ضرایب، کوچک بوده و تنها چند مؤلفه از سیگنال دارای مقادیر قابل توجهی باشند [5,6]. از این بحث در شاخه‌های مختلف علوم مهندسی مانند پردازش سیگنال گفتار و صوت، مهندسی دریا، آنتن، بازشناسی چهره، پردازش تصویر و غیره استفاده می‌شود. همان‌طورکه می‌دانیم بسیاری از سیگنال‌ها از جمله سیگنال گفتار در حوزه زمان نمایش تنک ندارند؛ اما با انتقال آن به فضای ویژگی جدید می‌توان احتمال نمایش تُنُک سیگنال را افزایش داد. به همین جهت تلاش‌های زیادی در راستای نمایش تُنُک این سیگنال‌ها در حوزه‌های مختلف صورت گرفته است. بسیاری از تبدیلاتی که در پردازش سیگنال مورد استفاده قرار می‌گیرند، واژه‌نامه‌های[2] کاملی را فراهم می‌آورند که هر یک برای نمایش دسته‌ای از سیگنال‌ها مناسب است. از جمله حوزه یا تبدیل‌های رایج می‌توان به تبدیل فوریه، تبدیل فوریه زمان کوتاه[3] (STFT)، تبدیل کسینوسی گسسته[4] (DCT) و تبدیل موجک[5] (WT) اشاره کرد. این نمایش‌های تُنُک، کاربرد زیادی در زمینه‌های مختلف همچون جداسازی منابع، نمایش تُنُک سیگنال، کدگذاری و غیره دارد. به همین جهت در سال‌های اخیر تلاش‌های زیادی برای ارائه نمایش‌های تُنُک این سیگنال‌ها صورت گرفته است. فضای ویژگی در مراجع معرفی‌شده حوزه بهسازی گفتار به‌طور عمده تاکنون شامل فرم اسپکتروگرام[6] بوده است. در این مقاله بهسازی گفتار به‌صورت تک‌کانال در حوزه تبدیل موجک با به‌کارگیری روش یادگیری واژه‌نامه و روال بازنمایی تُنُک صورت می‌پذیرد و از ضرایب حاصل از آن در سطح تجزیه موردنظر و در هر یک از زیرباندهای حاصل، به‌منظور آموزش و طراحی واژه‌نامه استفاده می‌شود. استفاده از این تبدیل به چند دلیل عمده مورد توجه قرار گرفته است: این تبدیل قابلیت سازگاری را با سیگنال‌های ناایستا مانند گفتار دارد. باند فرکانسی سیگنال ورودی جهت مدل‌کردن ویژگی

[1] Sparsity

[2] Dictionary
[3] Short Time Fourier Transform
[4] Discrete Cosine Transform
[5] Discrete wavelet Transform
[6] Spectrogram

[DOI: 10.29252/jsdp.17.3.17]

شنیداری گوش، باید به باندهای فرکانسی مختلف تجزیه شود. در این راستا، تبدیل موجک از این توانایی برخوردار بوده و می‌تواند به‌نحو مناسبی آنالیز فرکانسی انجام‌شده توسط گوش را با تجزیه سیگنال به زیرباندهای مختلف پیاده‌سازی کند. این تبدیل معکوس‌پذیر بوده و برگشت براساس آن به فضای داده آسان خواهد بود.

استفاده از تبدیل موجک و تحلیل زیرباندهای حاصل از آن یکی از زمینه‌های پردازشی مهم به‌منظور بهسازی سیگنال گفتار است [21-7]. در [7] یک الگوریتم آستانه‌گذاری برای ضرایب موجک ارائه شده و مقایسه نتایج بهسازی آن با سایر روال‌های آستانه‌گذاری مورد بررسی قرار گرفته است. در [8] یک روال بهسازی گفتار در فضای موجک با اعمال ضریب وزن‌دهی در هر زیرباند ارائه شده است. این ضرایب در طی روال بهسازی براساس نسبت واریانس سیگنال تخمینی به واریانس سیگنال نوفه‌ای تخمین زده می‌شوند. در [9]، از آستانه‌گذاری ضرایب تبدیل موجک در طراحی فیلتر وینر[1] در حوزه فرکانس به‌منظور حذف نوفهٔ ماشین از سیگنال گفتار استفاده شده است؛ همچنین یک الگوریتم حذف نوفهٔ گفتار در [10] براساس نوفهٔ سفید گوسی و توابع آستانه‌گذاری مختلف پیشنهاد شده است. در [11]، آستانه‌گذاری به‌روزشوندهٔ ضرایب موجک به‌کمک روال تجزیه حالت تجربی[2] و عمل‌گر انرژی Teager[3] در فضای تبدیل موجک به‌منظور حذف نوفه سفید گوسی به‌کار گرفته شده است. همچنین از این عمل‌گر انرژی در [12] نیز به‌منظور بهسازی سیگنال گفتار در فضای موجک استفاده شده است. همچنین در [17-13]، روش‌هایی به‌منظور بهسازی گفتار براساس ضرایب فضای تبدیل موجک ارائه شده است. در [21-18] نیز به بررسی نتایج بهسازی گفتار با استفاده از ضرایب بسته موجک[4] پرداخته می‌شود. تجزیه باند تقریب[5] در این تبدیل نیز به موازات تجزیه باند جزئیات[6] صورت می‌پذیرد. با توجه به پژوهش‌های انجام‌شده انتظار می‌رود که پردازش مناسب ضرایب این تبدیل، تخمین آستانه و اعمال آن به ضرایب مربوط به قاب‌های نوفه‌ای نتایج مناسبی در حوزه بهسازی گفتار به‌دست دهد.

در روش ارائه‌شده در این مقاله از مفاهیم بازنمایی تُنُک و یادگیری واژه‌نامه در جهت حذف انواع دسته نوفه‌های ایستا، ناایستا و متناوب از سیگنال گفتار استفاده می‌شود. روش پیشنهادی در دو سناریوی نظارت‌شده[7] و نیمه‌نظارت‌شده[8] پیشنهاد و ارائه می‌شود. در هر یک از این روال‌ها از آشکارسازی فعالیت گفتاری دقیقی به‌منظور تعیین قاب‌های نوفه‌ای جهت بازآموزی واژه‌نامه مربوط به داده نوفه بهره گرفته می‌شود. به‌روزرسانی اتم‌های مرتبط با سیگنال نوفه براساس به‌کارگیری روش تطبیق فضا[9] قاب‌های گفتاری و نوفه‌ای سیگنال مشاهده‌شده به‌کمک ضرایب موجک تعیین و برای هر یک از این دسته ضرایب، واژه‌نامه طراحی می‌شود [22,23]. از روش تطبیق فضا[10] به‌منظور دست‌یابی به واژه‌نامه متناسب با محیط آزمایش و شرایط نوفه‌ای آن بهره گرفته می‌شود [23]. همچنین یک روال بهینه‌سازی به‌منظور حصول اتم‌های با بیشینهٔ ناهمدوسی[11] به‌کمک انتقال فضای آموزش براساس داده نوفه‌ای ورودی مورد استفاده قرار می‌گیرد.

گام بازنمایی تُنُک در این روال نیز مبتنی‌بر یک الگوریتم طراحی‌شده به‌منظور دست‌یابی به اتم‌های با بیشترین همدوسی نسبت به قاب‌های داده درنظرگرفته‌شده خواهد بود. معیار مهمی که می‌بایست در یادگیری واژه‌نامه فراکامل[12] مورد توجه قرار گیرد، پارامتر همدوسی[13] میان اتم‌ها یا ستون‌های واژه‌نامه است. کوچکتربودن این معیار بیان‌گر این مسأله است که هر اتم مستقل از سایر اتم‌ها در نمایش قاب داده ورودی تأثیرگذار خواهد بود؛ بنابراین به‌کمک اتم‌های آموزش‌دیده تا حد ممکن مستقل، کمترین خطای تقریب در نمایش تُنُک سیگنال تصویر به‌دست می‌آید.

در بخش دوم این مقاله مسأله حذف نوفه از سیگنال گفتار مورد بررسی قرار می‌گیرد؛ سپس در بخش سوم، الگوریتم حذف نوفه پیشنهادی معرفی شده و در بخش چهارم نتایج شبیه‌سازی‌های انجام‌شده ارائه می‌شود. در ادامه نتیجه‌گیری در مورد روش معرفی‌شده بیان خواهد شد.

۲- تعریف مسأله

به‌طور معمول گام ابتدایی در یک الگوریتم بهسازی گفتار استفاده از یک حوزه تبدیل و انتقال سیگنال نوفه‌ای به یک فضای ویژگی جدید به‌منظور تحلیل مناسب‌تر است. این حوزه در روش پیشنهادی، فضای تبدیل موجک در نظر

[1] Wiener filter
[2] Empirical mode decomposition
[3] Teager energy operator
[4] Wavelet packet transform
[5] Approximation subband
[6] Detail subband
[7] Supervised
[8] Semi-supervised
[9] Domain transfer
[10] Domain adaptation
[11] Incoherency
[12] Overcomplete
[13] Coherence

[DOI: 10.29252/jsdp.17.3.17]

گرفته شده است. سیگنال گفتار نوفه‌ای در این فضا به‌صورت زیر مدل می‌شود:

$$Y(n) = S(n) + N(n) \quad (۱)$$

در این رابطه S و N به‌ترتیب سیگنال گفتار و نوفه تجزیه‌شده در هر زیرباند n است. ماتریس داده یک سیگنال $Y \in \mathbb{R}^{P\times F}$ می‌تواند با ترکیب خطی تُنُکی از اتم‌ها به‌صورت $Y=D.C$ که $D \in \mathbb{R}^{P\times L}, L > P$ یک واژه‌نامه فراکامل است، نمایش داده شود. واژه‌نامه شامل L اتم در ستون‌ها $\{d_l\}_{l=1}^{L}$ با نُرم واحد $\|d_{(:,l)}\|_2 = 1, \forall l = 1, \dots, L$ و بردار کدگذار K-تُنُک[1] $C \in \mathbb{R}^{L\times F}, L \gg K$ شامل ضرایب بازنمایی سیگنال Y خواهد بود [24,25]؛ بنابراین مسأله بازنمایی تُنُک که شامل بخش‌های خطای بازسازی و قید تُنُکی است، به‌صورت زیر قابل بیان است [24,25]:

$$C^* = arg\ \min_{C} \|Y - DC\|^2\ , \|C\|_0 \le K \quad (۲)$$

تعداد ضرایب غیرصفر در ماتریس ضرایب تُنُک C به‌صورت $\|C\|_0 = K$ بیان‌گر کارینالیتی سیگنال است. نُرم l_0 در رابطه (۲) به یک مسأله غیرمحدب منتهی می‌شود که NP-دشوار[2] است و ممکن است، موجب گرفتارشدن مسأله بهینه‌سازی در کمینه‌های محلی شود. آزادسازی[3] این مسأله با جایگزینی نُرم l_0 با نُرم l_1 در [26] گزارش شده است.

۳- روش پیشنهادی

روندنمای روش پیشنهادشده در این مقاله که شامل دو مرحله اساسی آموزش و آزمون است، در شکل (۱) نشان داده شده است. در گام نخست که آموزش واژه‌نامه‌ها صورت می‌گیرد، داده‌های آموزش با تبدیل موجک به فضای ویژگی انتقال پیدا می‌کنند. واژه‌نامه مرتبط با هر داده گفتار و نوفه براساس روالی که در ادامه معرفی می‌شود، آموزش داده می‌شود. توضیحات درخصوص هر یک از بلوک‌های این روند در ادامه آورده شده است.

۳-۱- بازنمایی تُنُک

آموزش مدل و بازنمایی تُنُک یکی از زمینه‌های پژوهشی جدید در حوزه‌های مختلف پردازش سیگنال است [27,28]. هدف از بازنمایی تُنُک، مدل‌کردن قاب‌های سیگنال ورودی به‌عنوان ترکیب خطی از تعداد کمی از بردارهای پایه است.

[1] K-sparse
[2] Non polynomial hard
[3] Relaxation

مسأله کدگذاری تُنُک داده آموزش بر روی واژه‌نامه مرکب به‌صورت زیر بیان می‌شود:

$$C_S^*, C_N^* = arg\ \min_{C_S, C_N} \|Y - DC\|_2 = arg\ \min_{C_S, C_N} \left\| Y - [D_S\ D_N] \begin{bmatrix} C_S \\ C_N \end{bmatrix} \right\|_2 \quad (۳)$$

ماتریس‌های C_S و C_N به‌ترتیب شامل ضرایب بازنمایی تُنُک متناظر با واژه‌نامه‌های گفتار و نوفه D_S و D_N است. در روش پیشنهادی از الگوریتم بازنمایی تُنُک LARC[4] استفاده شده که تعمیمی از الگوریتم LARS[5] با شرط توقف براساس مقدار همدوسی مانده[6] بوده و سعی به بیشینه‌کردن همدوسی متقابل در گام بازنمایی تُنُک را دارد [29,30]؛ از این‌رو رابطه (۳) می‌تواند به‌صورت زیر بیان شود:

$$C_S^*, C_N^* = LARC([D_S\ D_N], K, Coh) \quad (۴)$$

مقدار پارامتر Coh[7] در این رابطه بیان‌گر کمینه مقدار همدوسی میان اتم‌دادهٔ لازم برای پذیرش اتم به‌عنوان ستونی از واژه‌نامه و K نرخ تُنُکی است. ماتریس ضرایب تُنُک به‌دست‌آمده متناظر با وزن‌های واژه‌نامه‌های گفتار و نوفه، تخمینی از طیف گفتار $\hat{S}$ و طیف نوفه $\hat{N}$ را به‌دست می‌دهد:

$$\hat{S} = D_S C_S^*\ ,\ \hat{N} = D_N C_N^* \quad (۵)$$

درنظرگرفتن معیار همدوسی مانده برای خاتمه‌دادن به الگوریتم کدگذاری تُنُک، قابلیت مصالحه میان انحراف منبع و اعوجاج منبع را به‌دست می‌دهد. انحراف منبع زمانی رخ می‌دهد که بازنمایی تُنُک با نرخ تُنُکی پایین انجام یا بازنمایی بسیار تُنُک باشد؛ بنابراین تعداد اتم‌های مورد نیاز برای نمایش کافی نبوده و داده نمی‌تواند به‌درستی بر اتم‌های واژه‌نامه متناظر کد شود. اگر هر قاب تنها با واژه‌نامه متناظر خود بازنمایی شود، ERC[8] برقرار و کمینه انحراف رخ می‌دهد [29]:

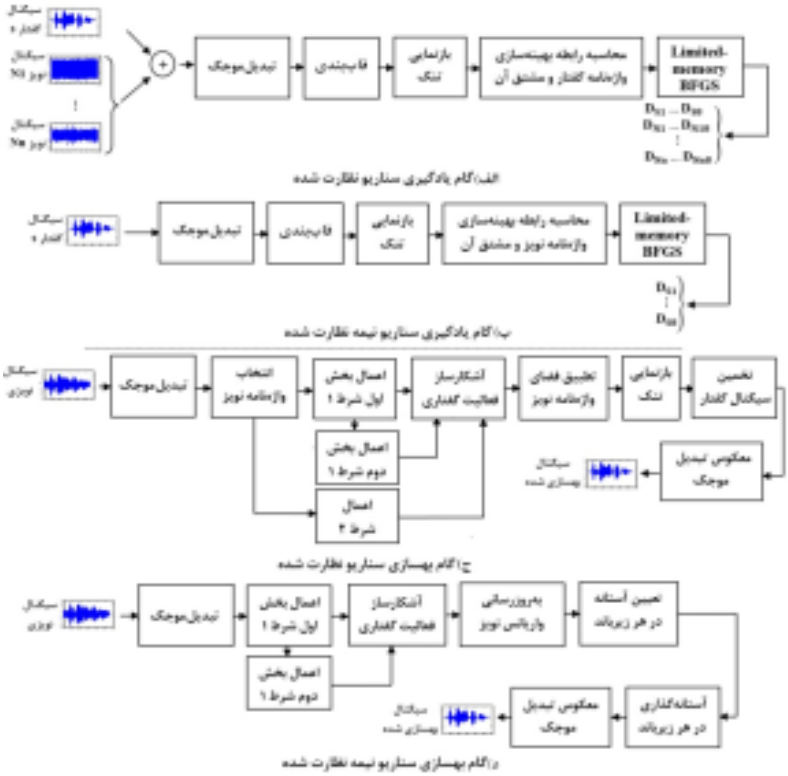
$$\|C_S^*\|_0 + \|C_N^*\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(D_S, D_N)}\right) \quad (۶)$$

در این رابطه، μ مقدار همدوسی میان دو واژه‌نامه گفتار و نوفه است. همچنین همدوسی متقابل میان اتم‌های واژه‌نامه گفتار d^S و نوفه d^N به‌صورت زیر به‌دست می‌آید [22]:

$$\mu(D_S, D_N) = \max_{1\le i\le L, 1\le j\le L} \left| d_i^S . d_j^N \right| \quad (۷)$$

[4] Least angle regression with coherence criterion
[5] Least-angle regression
[6] Residual coherence
[7] Coherence
[8] Exact recovery condition

[DOI: 10.29252/jsdp.17.3.17]



(شکل-۱): روندنمای روش پیشنهادی کاهش نوفه سیگنال گفتار مبتنی‌بر یادگیری واژه‌نامه در حوزه تبدیل موجک، شامل: الف) گام یادگیری در سناریوی نظارت‌شده. ب) گام یادگیری در سناریوی نیمه‌نظارت‌شده. ج) گام بهسازی در سناریوی نظارت‌شده. د) گام بهسازی در سناریوی نیمه‌نظارت‌شده.

(Figure-1): The block diagram of the proposed speech enhancement method based on dictionary learning technique in wavelet transform domain included: a) Training step in the supervised scenario. b) Training step in the semi supervised scenario. c) Enhancement procedure in the supervised scenario. d) Enhancement procedure in the semi supervised scenario.

از طرف دیگر، انتخاب پارامتر تُنُکی با نرخ بالا یا کدگذاری بسیار متراکم موجب اعوجاج منبع می‌شود؛ یعنی تعداد اتم‌ها در واژه‌نامه گفتار برای بازنمایی درست آن بسیار زیاد خواهد بود و از برخی اتم‌های واژه‌نامه نوفه نیز در بازنمایی سیگنال گفتار استفاده می‌شود؛ بنابراین انتخاب مقدار تُنُکی باید به‌دقت صورت پذیرد. در روش پیشنهادی از این روش بازنمایی تُنُک در گام نخست الگوریتم یادگیری واژه‌نامه به‌منظور دست‌یابی به واژه‌نامه‌های با همدوسی متقابل بالا استفاده شده است. با توجه به شکل (۱)، روش بازنمایی تُنُک LARC[1] بر روی ماتریس داده در هر زیرباند اعمال و ماتریس ضرایب تُنُک C_S و C_N متناظر با هر یک از زیرباندها حاصل می‌شوند [29].

۲-۳- روال یادگیری واژه‌نامه

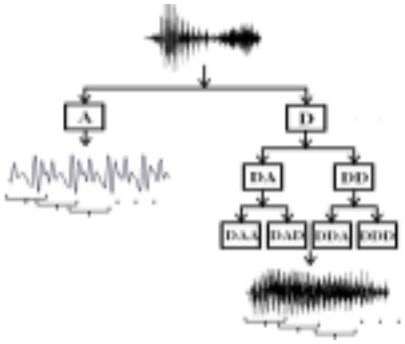
در گام آموزش، در ابتدا برای هر یک از سیگنال‌های گفتار و داده‌های نوفه، ضرایب موجک با سطح تجزیه سه در باند جزئیات[2] و یک سطح تجزیه در باند تقریب[3] محاسبه و پنج زیرباند متناظر برای هر قاب داده ورودی ایجاد می‌شود. داده حاصل در هر یک از پنج باند، قاب‌بندی و برای قاب‌های حاصل از هر زیرباند، یادگیری واژه‌نامه به‌صورت مجزا انجام می‌شود. این روال پیش‌پردازش در شکل (۲) نشان داده شده است. بلوک A و D در این شکل به‌ترتیب مربوط به زیرباند تقریب و جزئیات است. به این‌صورت برای هر داده ورودی پنج واژه‌نامه مرتبط با پنج زیرباند به‌دست می‌آید. گفتار یک سیگنال با ناایستایی یا تغییرات بالا است؛ بنابراین تحلیل مولفه‌های فرکانس بالا در آن بسیار حائز اهمیت است. همچنین سیگنال‌های نوفهٔ مورد بررسی در این مقاله نیز به جز نویز سفید و موسیقی دارای ناایستایی بالا هستند که از این جهت حذف آن‌ها از سیگنال گفتار می‌بایست با دقت بیشتر انجام شود. به‌منظور بررسی دقیق مؤلفه‌های فرکانس بالا که تحت تأثیر شدید نوفه قرار می‌گیرند، تحلیل زیرباندهای جزئیات از اهمیت ویژه‌ای نسبت به باند تقریب

[1] Least angle regression with coherence criterion

[2] Detail
[3] Approximation

[DOI: 10.29252/jsdp.17.3.17]

برخوردار است؛ همچنین تجزیه بیشتر زیرباند تقریب اطلاعات بیشتری از سیگنال به‌دست نمی‌دهد.



(شکل-۲): قاب‌بندی داده هر زیرباند تبدیل موجک برای تجزیه سطح سه در باند جزئیات و سطح یک در باند تقریب.

(Figure-2): Input data framing in the last subband of approximation and detail decomposition level.

K-SVD یک روش یادگیری مؤثر برای دست‌یابی به واژه‌نامه فراکامل از مجموعه‌ای از سیگنال‌های نمونه است [31]. هر قاب ورودی در این الگوریتم با ترکیب خطی از ضرایب K اتم در یک روش مبتنی‌بر تجزیه مقادیر منفرد، بازنمایی می‌شود. دو ویژگی اساسی در یادگیری واژه‌نامه، همدوسی متقابل و همدوسی بین اتم‌ها است که همدوسی در اتم‌های یک واژه‌نامه و همدوسی اتم‌ها به داده آموزش را تعیین می‌کند. این پارامترها وقتی سیگنال‌های داده آموزش و نوفه‌های مورد نظر ساختار یکسان دارند باید بسیار مورد توجه قرار گیرد. همدوسی بالاتر با ردهٔ داده و همدوسی متقابل پایین‌تر میان اتم‌ها موجب می‌شود که خطای تقریب تُنُک کوچک‌تری حاصل شود (رابطه ۷). واژه‌نامه‌ها باید بیشترین میزان همدوسی متقابل را داشته باشند تا مانع از اعوجاج در سیگنال بازسازی‌شده شوند. این بدان معنی است که واژه‌نامه گفتار باید نسبت به هر یک از واژه‌نامه‌های نویز ناهمدوس باشد تا قاب‌های مشاهده‌شده از این سیگنال به‌درستی بر روی واژه‌نامه متناظر، کدگذاری شود. این معیار به‌صورت زیر محدود می‌شود:

$$\mu(D) \geq \sqrt{(L-P)/P(L-1)} \qquad (۸)$$

در این رابطه L تعداد اتم‌ها و P بعد هر اتم است. همچنین، شرط بازسازی درست[1]ERC، بازسازی مطلوب داده را تضمین می‌کند و منجر به نرخ سریع‌تر کاهش خطای مانده در هم‌گرایی الگوریتم یادگیری می‌شود [32]:

$$\|C\|_0 < \frac{1}{2}\left(1+\frac{1}{\mu(D)}\right) \qquad (۹)$$

[1] Exact recovery condition

در این رابطه $\|C\|_0$ تعیین کننده مقدار کاردینالیتی با نرخ بازنمایی تُنُک است. مسأله یافتن یک واژه‌نامه با همدوسی متقابل کم می‌تواند به مقیدکردن ماتریس گرام[2] واژه‌نامه $G = D^T D$ منتهی شود. مقدار همدوسی واژه‌نامه D بیشینه مقدار مطلق عناصر غیرقطری ماتریس گرام وقتی اتم‌ها نرمالیزه باشند، معرفی می‌شود [33].

این واژه‌نامهٔ نرمالیزه‌شده که به‌منظور بازسازی تُنُک مناسب مورد استفاده قرار می‌گیرد، قاب سخت متساوی‌الزاویه[3]ETF یا قاب گرسمانیان[4] نامیده می‌شود [34]؛ اما مسأله اساسی این است که این ماتریس برای هر ابعاد دلخواه از واژه‌نامه وجود ندارد؛ بنابراین حل مسأله به‌صورت تقریبی انجام خواهد گرفت. یکی از نزدیک‌ترین راه‌حل‌های قابل قبول، پس‌پردازش واژه‌نامه مورد نظر خواهد بود. هدف در روش‌های تقریبی مانند روش به‌کارگرفته‌شده در این مقاله، کاهش مقدار خطای تقریب $\|G-I\|_F$ برای دست‌یابی به ماتریس با کمینه همدوسی میان اتم‌ها خواهد بود.

از آنجایی‌که روش بازنمایی تُنُک OMP[5] انعطاف‌پذیر است از نتایج حاصل از آن همراه با هر روش یادگیری واژه‌نامه می‌توان استفاده کرد. روال یادگیری واژه‌نامه استفاده‌شده در الگوریتم پیشنهادی مبتنی‌بر روش Limited-memory BFGS[6] است که یک الگوریتم تکرارشونده برای حل مسائل بهینه‌سازی غیرخطی بدون قید خواهد بود و روش نیوتن را در حل مسأله تقریب می‌زند [35]. نحوه تخصیص داده آموزش در هر زیرباند به الگوریتم یادگیری واژه‌نامه، در شکل (۳) نمایش داده شده است. مسأله بهینه‌سازی در یادگیری واژه‌نامه با اتم‌های ناهمدوس برای داده گفتاری به‌صورت زیر خواهد بود:

$$F_{D_S} = arg\min_{D_S} \|S - D_S C_S\|_F^2 + \gamma \left\| D_S{}' D_S - I \right\|_F^2 \qquad (۱۰)$$

که بخش نخست مربوط به کاهش خطای تقریب داده گفتار و بخش دوم مربوط به یافتن واژه‌نامه با ماتریس گرام نزدیک به ETF که از پیش در مورد آن صحبت شده خواهد بود و γ پارامتر کنترل‌کننده مصالحه میان این دو بخش است. با درنظرگرفتن رابطه کلی $\|A\|_F^2 = tr\{A'A\}$، از رابطه (۱۰) خواهیم داشت:

[2] Gram matrix
[3] Equiangular tight frame
[4] Grassmannian frame
[5] Orthogonal matching pursuit
[6] Broyden-Fletcher-Goldfarb-Shanno

[DOI: 10.29252/jsdp.17.3.17]

[DOI: 10.29252/jsdp.17.3.17]

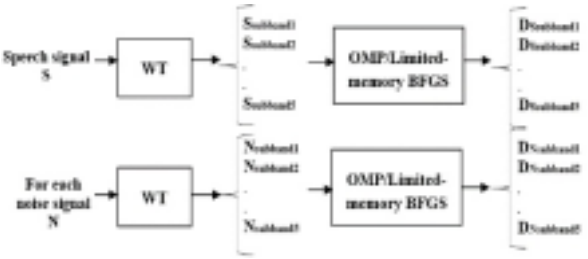
$$F_{D_S} = tr\{C_S'D_S'D_SC_S\} - 2tr\{S'D_SC_S\} + tr\{S'S\} + \gamma(tr\{D_S'D_SD_S'D_S\} - 2tr\{D_S'D_S\} + tr\{I\}) \quad (۱۱)$$

که سه بخش نخست مربوط به خطای تقریب و مابقی بخش‌ها مربوط به ماتریس گرام ETFاست. به‌منظور حل این مسأله به‌کمک روش Limited-memory BFGS، می‌بایست مشتق رابطه (۱۰) محاسبه شود:

$$\partial F_{D_S}/\partial D_S = 2(D_SC_SC_S' - SC_S') + 4\gamma(D_SD_S'D_S - D_S) \quad (۱۲)$$

مسأله بهینه‌سازی یادگیری واژه‌نامه نوفه با اتم‌های ناهمدوس نسبت به اتم‌های یادگیری‌شده برای داده گفتاری در مرحله قبل، به‌صورت زیر بیان می‌شود:

$$F_{D_N} = arg\min_{D_N}\|N - D_NC_N\|_F^2 + \gamma_1\|D_N'D_N - I\|_F^2 + \gamma_2\|D_N'D_S\|_F^2 \quad (۱۳)$$



(شکل-۳): نحوه تخصیص داده آموزش به‌منظور بهسازی گفتار در فضای ویژگی تبدیل موجک.

(Figure-3): Training process of speech and noise signals in different subbands of wavelet transform.

بخش نخست در این رابطه مربوط به کاهش خطای تقریب داده نوفه، بخش دوم مربوط به یافتن واژه‌نامه نوفه با ماتریس گرام نزدیک به ETF و بخش سوم مربوط با یافتن اتم‌های واژه‌نامه نوفهٔ مستقل از اتم‌های واژه‌نامه گفتار است. پارامتر γ_1 و γ_2 مصالحه میان این بخش‌ها را کنترل می‌کند. با ساده‌کردن این رابطه خواهیم داشت:

$$F_{D_N} = tr\{C_N'D_N'D_NC_N\} - 2tr\{N'D_NC_N\} + tr\{N'N\} + \gamma_1(tr\{D_N'D_ND_N'D_N\} - 2tr\{D_N'D_N\}) + tr\{I\} + \gamma_2(tr\{D_S'D_ND_N'D_S\}) \quad (۱۴)$$

در این رابطه سه بخش نخست مربوط به خطای تقریب، سه بخش دوم مربوط به بخش ماتریس گرام با ناهمدوسی بالا و بخش آخر مربوط به ناهمدوسی متقابل میان واژه‌نامه گفتار و نوفه است. مشتق رابطه (۱۳) به‌منظور حل با استفاده از روش Limited-memory BFGS به‌صورت زیر است:

$$\partial F_{D_N}/\partial D_N = 2(D_NC_NC_N' - NC_N') + 4\gamma_1(D_ND_N'D_N - D_N) + 2\gamma_2(D_N'D_SD_S') \quad (۱۵)$$

تابع F_{D_S} و F_{D_N} و مشتقات به‌دست‌آمده آن‌ها به‌عنوان ورودی به الگوریتم بهینه‌سازی یادشده اعمال و واژه‌نامه گفتار به‌همراه واژه‌نامه برای هر نوع نوفه با نرخ ناهمدوسی بالا به‌دست می‌آید. گفتنی است که در سناریوی بهسازی گفتار نظارت‌شده، هر دو مرحله یادگیری واژه‌نامه برای گفتار و نوفه و در سناریوی نیمه‌نظارت‌شده تنها گام یادگیری واژه‌نامه گفتار انجام می‌شود.

۳-۳- انتخاب واژه‌نامه نوفه

سیگنال نوفه‌ای در گام بهسازی پیشنهادی در ابتدا به فضای تبدیل موجک انتقال می‌یابد و سپس نوع نوفه‌ای که گفتار با آن مواجه شده تشخیص داده می‌شود؛ ازاین‌رو واژه‌نامه مرکب تنها شامل دو واژه‌نامه گفتار و نوفه به جای واژه‌نامه مرکب متشکل از واژه‌نامه گفتار و تمامی واژه‌نامه‌های نوفه خواهد بود. این مسأله موجب می‌شود که نتایج در بازنمایی تُنُک گام بهسازی با سرعت بیشتری رخ دهد و با مشکلات پیچیدگی محاسبات مواجه نشویم. در [29]، الگوریتم بهسازی گفتار دیگری با استفاده از یادگیری واژه‌نامه مطرح شده است که در آن اتم‌های واژه‌نامه گفتار و نوفه‌های مورد بررسی به‌صورت برون‌خط آموزش می‌بینند. داده نوفهٔ مورد استفاده از نوفه خالص بوده که به‌طور معمول نوفه شامل در سیگنال نوفه‌ای بر روی آن بازنمایی تُنُک خیلی دقیق نخواهد داشت؛ همچنین واژه‌نامه مرکب در گام بهسازی متشکل از واژه‌نامه گفتار و تمامی واژه‌نامه‌های نوفهٔ آموزش‌دیده است. این مسأله ممکن است، عملکرد الگوریتم در زمینه به‌دست‌آوردن نتایج بهسازی مناسب را با مشکل مواجه کند و منجر به کدگذاری خیلی متراکم سیگنال گفتار یا اغتشاش گفتار[1] شود؛ زیرا ممکن است، برخی قاب‌های گفتاری با استفاده از اتم‌های نوفه مختلف به‌صورت تُنُک بازنمایی شوند؛ همچنین زمان محاسباتی الگوریتم در گام بهسازی به‌علت ابعاد بالای واژه‌نامه مرکب بسیار بالا خواهد بود. مسأله دیگر این است که ساختار مؤلفه‌های برخی قاب‌ها در سیگنال‌های نوفه مختلف به‌صورت ذاتی مشابه هستند. ازاین‌رو در واژه‌نامه مرکب حاصل، اتم‌های مشابه وجود

[1] Speech confusion

خواهند داشت و مؤلفۀ نوفه سیگنال مشاهده‌ای به‌درستی نمی‌تواند بر روی واژه‌نامه نوفه متناظر بازنمایی تُنُک داشته باشد. به‌منظور انتخاب نوع نوفه، قاب‌های اولیه سکوت گفتار به‌صورت تُنُک بر روی واژه‌نامه مرکبی که شامل تمامی واژه‌نامه‌های نوفه است، بازنمایی داده می‌شود؛ سپس انرژی ضرایب تُنُک در کدگذاری بر روی هر واژه‌نامه محاسبه می‌شود. نوع نوفۀ مشاهده‌شده با توجه به واژه‌نامه نوفه با بیشترین انرژی ضرایب بازنمایی حاصل می‌شود؛ زیرا هر ردۀ داده تنها بر روی واژه‌نامه ناهمدوس طراحی‌شدۀ متناظر خود بازنمایی خواهد داشت. ازاین‌رو، واژه‌نامه مرکب حاصل از واژه‌نامه گفتار و تنها واژه‌نامه نوفه انتخابی، زمان محاسباتی برای اجرای روال بهسازی را به‌طور چشم‌گیری کاهش می‌دهد.

۴-۳- الگوریتم بهسازی سیگنال گفتار

در نخستین مرحله از گام بهسازی، برچسب قاب‌های گفتار/غیرگفتار به‌کمک الگوریتمی که در ادامه معرفی می‌شود، تعیین می‌شود. در این روال که برای هر زیرباند و در هر قاب انجام می‌شود، در ابتدا و در شرط نخست شباهت میان قاب داده و قاب دادۀ بازسازی‌شده که به‌کمک ضرایب تُنُک حاصل از بازنمایی به روی واژه‌نامه گفتار D_S به‌دست می‌آید، بررسی می‌شود. اگر شباهت میان این دو قاب زیاد باشد؛ به معنای آن است که قاب ورودی گفتاری بوده و به همین دلیل بر روی واژه‌نامه گفتار بازنمایی مناسب داشته است؛ بنابراین به‌منظور اطمینان از تصمیم گرفته‌شده، انرژی ضرایب تُنُک حاصل از بازنمایی تُنُک قاب مشاهده‌شده بر روی قاب‌های ابتدایی و انتهایی سیگنال گفتار که ساختار نوفه‌ای دارند و در روابط با D_{N0} مشخص شده‌اند، بررسی می‌شود. اگر انرژی ضرایب تُنُک کم باشد، به معنای این است که قاب ورودی به‌علت دارابودن ماهیت گفتاری بر روی اتم‌های با ماهیت نوفه‌ای، بازنمایی مناسب نداشته و تصمیم گرفته‌شده در مورد گفتاری‌بودن قاب مورد بررسی درست بوده است. از طرف دیگر اگر شباهت میان این دو قاب کم باشد که به معنای خطای بازنمایی زیاد قاب ورودی بر روی واژه‌نامه گفتار است، می‌توان نتیجه گرفت که قاب غیرگفتاری است. بنابراین به‌منظور اطمینان بیشتر به تصمیم گرفته‌شده، قاب ورودی بر روی قاب‌های داده ابتدایی و انتهایی سیگنال گفتار D_{N0} که ماهیت نوفه‌ای دارند، بازنمایی می‌شوند و اگر انرژی ضرایب تُنُک در این بازنمایی بالا باشد، به معنای اطمینان به تصمیم گرفته‌شده مبنی بر ماهیت غیرگفتاری‌داشتن قاب ورودی است [36].

این شرط مخصوص سناریوی نیمه‌نظارت‌شده است و اگر سناریو نظارت‌شده باشد شرط دیگری نیز برای اطمینان به تصمیم گرفته‌شده وجود دارد که به‌منظور تصمیم‌گیری مناسب در کنار شرط نخست قرار می‌گیرد. شرط دوم با توجه به دراختیاربودن واژه‌نامه نوفه علاوه‌بر واژه‌نامه گفتار درنظر گرفته می‌شود. در شرط دوم انرژی ماتریس ضرایب تُنُک که حاصل از بازنمایی قاب داده ورودی بر روی واژه‌نامه مرکب است، بررسی می‌شود. واژه‌نامه مرکب همان‌طور که از پیش بیان شده از کنار هم قرارگرفتن واژه‌نامه گفتار و واژه‌نامه نوفه انتخاب‌شده به‌دست می‌آید. از مقایسه انرژی متناظر با هر واژه‌نامه می‌توان تصمیم مناسبی درخصوص ماهیت قاب ورودی اتخاذ کرد. اگر قاب ورودی گفتاری باشد، انرژی ضرایب تُنُک آن بر روی واژه‌نامه گفتار بسیار بالاتر از انرژی ضرایب تُنُک حاصل بر روی واژه‌نامه نوفه خواهد بود. اگر این شرط در هر چهار زیرباند تقریب برقرار باشد، برچسب گفتاری از شرط دوم به قاب ورودی اختصاص می‌یابد. از آنجایی‌که امکان شباهت در زیرباند جزئیات برای داده گفتار و نوفه زیاد است، تأثیر مقایسه انرژی ضرایب این مجموعه کم‌تر خواهد بود [36].

بررسی شرط نخست: تعیین میزان شباهت میان قاب داده و قاب داده بازسازی شده:

$$C = OMP(Y, D_S) \rightarrow \hat{Y} = D_S.C \rightarrow < Y, \hat{Y} > = Y.\hat{Y} \Big/ \left(\|Y\|_2 . \|\hat{Y}\|_2\right)$$

۱- اگر $\|Y - \hat{Y}\|_2 < \varepsilon$ که به معنای گفتاری‌بودن قاب مشاهده‌شده است، بازنمایی بر روی قاب‌های اولیه نوفه‌ای صورت می‌گیرد:

$$C_0 = LARC(Y, D_{N0}) \rightarrow {}^1\!/_L \sum_{l=1}^{L} \mathrm{C}_0^2$$

اگر انرژی ضرایب تُنُک در این بازنمایی کم باشد به معنای دارابودن برچسب گفتاری برای قاب ورودی است. L در این رابطه طول قاب مورد بررسی خواهد بود.

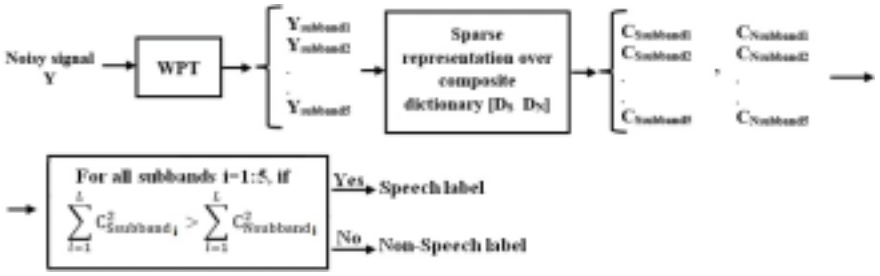
۲- به همین صورت اگر $\|Y - \hat{Y}\|_2 > \varepsilon$ که به معنای نوفه‌ای‌بودن قاب مشاهده‌شده است، بازنمایی بر روی قاب‌های اولیه نوفه‌ای صورت می‌گیرد؛ اگر انرژی ضرایب تُنُک در این بازنمایی زیاد باشد، به معنای اطلاق برچسب غیرگفتاری به قاب ورودی است.

[DOI: 10.29252/jsdp.17.3.17]

بررسی شرط دوم: در ابتدا بازنمایی تُنُک بر روی واژه‌نامه مرکب انجام و سپس انرژی ضرایب متناظر با هر واژه‌نامه بر روی قاب محاسبه و مقایسه بر روی هر چهار زیرباند تقریب و جزئیات صورت می‌گیرد. چگونگی عملکرد این شرط در آشکارساز فعالیت گفتاری پیشنهادی برای هر زیرباند تبدیل موجک در شکل (۴) نشان داده شده است؛ درنهایت می‌توان با توجه به نتایج حاصل از شرط نخست و دوم در مورد اختصاص برچسب گفتاری یا غیرگفتاری به هر قاب ورودی تصمیم‌گیری کرد. در گام بهسازی با توجه به نتایج به‌دست‌آمده در مرحله قبل و تشخیص قاب‌های گفتار/غیرگفتار، با توجه به سناریوی مورد بررسی دو راه‌کار پیشنهاد شده است. در سناریوی نظارت‌شده، به‌کمک قاب‌های نوفه‌ای به‌دست‌آمده از نتایج مرحله قبل و تکنیک تطبیق فضا، اتم‌های واژه‌نامه نوفه از پیش یادگیری‌شده متناسب با شرایط نوفه محیط آزمون و به‌منظور انطباق بیشتر با آن به‌روز می‌شوند؛ که این واژه‌نامه تطبیق‌یافته با $\widehat{D}_N$ معرفی می‌شود؛ سپس بازنمایی تُنُک قاب ورودی بر روی واژه‌نامه مرکب حاصل از واژه‌نامه گفتار و این واژه‌نامه تطبیق‌یافته انجام می‌گیرد [36]:

$$Y = [D_S\ \widehat{D}_N]\begin{bmatrix} C_S \\ \hat{C}_N \end{bmatrix} \rightarrow \hat{S} = D_S.C_S \qquad (۱۶)$$

در سناریوی نیمه‌نظارت‌شده که واژه‌نامه نوفه در دسترس نیست، به‌کمک روابط به‌روزکردن تخمین نوفه در هر قاب و زیرباند، رابطه آستانه‌گذاری ضرایب موجک متناسب با سطح نوفه‌ای که قاب درگیر آن است، تغییر می‌کند. به‌طور معمول آستانه‌گذاری ضرایب در تمامی سطوح و زیرباندها با نرخ ثابتی صورت می‌گیرد و صرف‌نظر از سطح نوفه‌ای است که هر زیرباند را آلوده کرده است. این مسأله موجب عدم دست‌یابی به نتایج مطلوب در فرآیند بهسازی می‌شود؛ بنابراین در روش پیشنهادی با توجه به دانسته‌های حاصل در مورد قاب‌های نوفه‌ای می‌توان واریانس نوفه را در هر قاب به‌روز کرده و حذف ضرایب با توجه به آن در هر زیرباند صورت گیرد. همان‌طورکه بیان شد روندنمای کلی روش پیشنهادی در حوزه تبدیل موجک به‌تفکیک دو سناریوی نظارت‌شده و نیمه نظارت‌شده مورد بحث در شکل (۱) نمایش داده شده است.



(شکل-۴): نحوه عملکرد آشکارساز فعالیت گفتاری پیشنهادشده در فضای تبدیل موجک در هر زیرباند با درنظرگرفتن شرط دوم.

(Figure-4): The procedure of label detection in the proposed energy-based VAD algorithm in different subbands of wavelet decomposition.

باید توجه داشت که هر یک از این مراحل برای هر زیرباند از پنج زیرباند مورد بررسی انجام می‌گیرد. رابطه به‌روزرسانی سطح نوفه در هر قاب به‌صورت زیر خواهد بود:

$$\hat{N}_{j,n} = \alpha\hat{N}_{j,n-1} + (1-\alpha)Y_{j,n}^2 \qquad (۱۷)$$

j شماره زیرباند و n شماره قاب است. انحراف معیار نویز در هر زیرباند به‌صورت زیر در نظر گرفته می‌شود [35]:

$$\hat{\sigma}_j = {median\left(|\hat{N}_j|\right)}/{0.6745} \qquad (۱۸)$$

عبارت $median$ بیانگر انحراف مطلق میانه[1] است. محاسبه مقدار آستانه در هر زیرباند به‌صورت زیر است [37]:

$$T_j = \hat{\sigma}_j\sqrt{2ln(L_j)} \qquad (۱۹)$$

در این رابطه، طول قاب در زیرباند j ام است. L_j

همچنین تابع آستانه‌گذار به‌صورت زیر خواهد بود:

$$\hat{Y} = \begin{cases} Y - \frac{1}{2}\frac{T_j^2}{Y}, & |Y| \geq T_j \\ \frac{1}{2}sign(Y).\frac{|Y|^2}{T_j}, & |Y| < T_j \end{cases} \qquad (۲۰)$$

در این رابطه شیب حذف ضرایب در تابع آستانه‌گذار متناسب با سطح نویز تغییر می‌کند. هر چقدر مقدار پارامتر آستانه T_j بیشتر باشد به این معنا است که نوفه بیشتری می‌بایست از سیگنال حذف شود پس شیب منحنی در بخش $|Y| < T_j$ کاهش پیدا کرده و امکان حذف ضرایب بیشتری در هر زیرباند فراهم می‌شود. درنهایت ضرایب آستانه‌گذاری‌شده حاصل با تبدیل معکوس موجک به فضای داده برگشت داده می‌شوند.

[1] Median absolute deviation

[DOI: 10.29252/jsdp.17.3.17]

(جدول-۱): مقادیر همدوسی متقابل میان اتم‌ها و SNR تقریب سیگنال داده برای واژه‌نامه گفتار و نوفه بر حسب dB در روش‌های مختلف.

(Table -1): The atom coherence and SNR values in signal reconstruction for speech and noise dictionaries using different methods.

		گفتار	همهمه	کارخانه	سفید	ماشین	خیابان	قطار	رستوران	پیانو
K-SVD /OMP	همدوسی	۰/۸۷	۰/۸۹	۰/۹۱	۰/۹۲	۰/۸۹	۰/۹۰	۰/۸۸	۰/۸۹	۰/۹۴
	SNR	۱۲/۴	۱۰/۶	۱۰/۸	۱۱/۲	۱۲/۱	۱۲	۱۰/۷	۱۰/۸	۱۱/۱
K-SVD /LARC	همدوسی	۰/۷۶	۰/۸۳	۰/۹۰	۰/۹۵	۰/۹۴	۰/۹۲	۰/۸۹	۰/۹۱	۰/۹۶
	SNR	۱۲/۴	۱۰/۶	۱۰/۸	۱۱/۱	۱۱/۷	۱۱/۸	۱۰/۴	۱۰/۷	۱۱/۲
روش پیشنهادی (رابطه ۹)	**همدوسی**	**۰/۳۷**	**۰/۴۹**	**۰/۴۱**	**۰/۴۶**	**۰/۴۴**	**۰/۴۵**	**۰/۴۶**	**۰/۴۲**	**۰/۵۰**
	SNR	**۱۲/۵**	**۱۱/۲**	**۱۱/۳**	**۱۱/۶**	**۱۲/۳**	**۱۲/۲**	**۱۱**	**۱۱/۳**	**۱۱/۵**

(جدول-۲): مقادیر همدوسی متقابل میان اتم‌های واژه‌نامه گفتار و هر یک از واژه‌نامه‌های نوفه در روش‌های مختلف.

(Table -2): The mutual coherence between speech and each of noise dictionary atoms using different methods.

	همهمه	کارخانه	سفید	ماشین	خیابان	قطار	رستوران	پیانو
K-SVD/OMP	۰/۹۳	۰/۸۵	۰/۷۸	۰/۸۶	۰/۸۹	۰/۹۰	۰/۹۱	۰/۷۷
K-SVD/LARC	۰/۹۱	۰/۸۷	۰/۷۳	۰/۸۸	۰/۹۱	۰/۹۰	۰/۹۳	۰/۷۲
روش پیشنهادی (رابطه ۱۲)	**۰/۶۸**	**۰/۶۰**	**۰/۵۷**	**۰/۶۳**	**۰/۶۱**	**۰/۵۹**	**۰/۶۳**	**۰/۵۶**

۴- جزئیات شبیه‌سازی

شبیه‌سازی‌های انجام‌شده بر روی مجموعه‌دادگان جامع گفتاری GRID با تعداد گوینده و عبارت بیان‌شده به تعداد زیاد انجام گرفته که مناسب برای شبیه‌سازی‌های این مقاله در وضعیت وابسته به گوینده (SD)[1] و مستقل از گوینده (SI)[2] خواهد بود [38]. همچنین نرخ نمونه‌برداری برابر با ۸kHz در نظر گرفته شده است. گام آزمون در این مقاله در دو وضعیت مستقل از گوینده (گوینده‌های متفاوت در مرحله آزمون و آموزش) و وابسته به گوینده (گوینده‌های مشابه در مرحله آزمون و آموزش) اجرا می‌شود. مجموعه آموزش و آزمون به‌ترتیب شامل صدوپنجاه عبارت گفتاری خواهد بود. از این میان هشت گوینده زن و هشت گوینده مرد در گام آموزش و از سه گوینده مرد و سه گوینده زن در آزمون مستقل از گوینده استفاده شده است. طول قاب داده برابر با ۲۰ms و قاب‌ها با هم‌پوشانی ٪۵۰ تنظیم شده‌اند. تنظیمات در روال یادگیری برای همه منابع داده شامل گفتار و نوفه‌های مختلف یکسان است. نرخ افزونگی واژه‌نامه گفتار برابر چهار و نرخ افزونگی واژه‌نامه نوفه در سناریوی نظارت‌شده نیز برابر همین مقدار در نظر گرفته شده است. شبیه‌سازی بر روی محدوده گسترده‌ای از سیگنال‌های نوفه شامل نوفه همهمه(Bab)، کارخانه(Fct)، سفید(Wht)، ماشین(Car)، خیابان(Str)، قطار(Trn)، رستوران(Rst) و نوفه موسیقی پیانو(Pno) از وب‌سایت جامعه پیانو صورت گرفته است [39-41]؛ همچنین با توجه به بررسی‌های انجام‌شده، سطح تجزیه باند جزئیات برابر سه درنظر گرفته شده است. مقادیر آستانه ε در شرط نخست و دوم روال آشکارسازی فعالیت گفتاری پیشنهادی براساس نتایج شبیه‌سازی‌ها و به‌صورت تجربی تنظیم شده‌اند. این مقدار برای تمامی نوفه‌ها، مقدار ۰/۰۱۲۵ در نظر گرفته شده است؛ همچنین پارامتر α براساس نتایج تجربی و برای تمامی نوفه‌ها به مقدار ۰/۸۵ تنظیم شده است. در نخستین شبیه‌سازی انجام‌شده به‌منظور ارزیابی عملکرد در روال یادگیری واژه‌نامه ارائه‌شده، دو پارامتر همدوسی متقابل اتم‌ها و همدوسی اتم‌های واژه‌نامه‌های مختلف مورد بررسی قرار گرفت. مقدار همدوسی متقابل اتم‌ها در واژه‌نامه گفتار و واژه‌نامه‌های نوفه به‌دست‌آمده از رابطه (۷) برای روش‌های یادگیری واژه‌نامه براساس الگوریتم K-SVD با گام بازنمایی تُنُک OMP، الگوریتم K-SVD با گام بازنمایی تُنُک LARC، الگوریتم K-SVD با گام بازنمایی تُنُک LARC دنبال‌شده با روش IPR و روش استفاده‌شده در این مقاله براساس Limited memory-BFGS، مورد بررسی قرار گرفته است. نتایج این شبیه‌سازی در جدول (۱) نشان داده شده است. در این جدول، مقدار SNR گزارش‌شده از رابطه زیر برای هر واژه‌نامه محاسبه شده است:

$$SNR(Y, DC) = 20\, log_{10}(\|Y\|_F / \|Y - DC\|_F) \qquad (۲۱)$$

در این رابطه Y، سیگنال تقریب زده‌شده در بازنمایی تُنُک است. همان‌طورکه پیش از این بیان شد در روش پایه یادگیری واژه‌نامه مبتنی‌بر K-SVD/OMP هیچ‌گونه روش کاهش همدوسی متقابل اتم‌ها به‌کارگرفته نشده است. به

[1] Speaker dependent
[2] Speaker independent

[DOI: 10.29252/jsdp.17.3.17]

همین دلایل مقادیر همدوسی گزارش‌شده برای این روش بالا خواهد بود. همچنین در روال یادگیری مبتنی‌بر K-SVD/LARC تنها پارامتر همدوسی اتم‌ها به داده آموزش بررسی می‌شود و سعی بر این است که بیشینه همدوسی بین اتم‌ها و رده داده مورد نظر برقرار شود [29]. بنابراین مقدار SNR به‌دست‌آمده در این روش بالاتر از روال یادگیری K-SVD/OMP خواهد بود؛ همچنین، در روال یادگیری K-SVD/LARC/IPR که همان روش پیشنهادی است، از گام پس‌پردازش IPR به‌منظور کاهش همدوسی متقابل اتم‌های واژه‌نامه طی دو مرحله و به‌منظور نزدیک‌کردن ماتریس گرام واژه‌نامه به ساختار ETF استفاده شده، [34] به همین دلیل مقادیر همدوسی متقابل اتم‌ها در این روش یادگیری کوچک‌تر از دو مرحله یادشدهٔ قبلی است. در روش به‌کارگرفته‌شده در این مقاله، بخشی از تابع بهینه‌سازی مورد استفاده برای یادگیری واژه‌نامه براساس روابط (۱۰ و ۱۳) به‌منظور یافتن ساختار ETF برای واژه‌نامه درنظر گرفته شده است و بهینه‌سازی اتم‌ها بر این مبنا و همراه با کاهش خطای تقریب داده صورت می‌پذیرد. همان‌طورکه دیده می‌شود، روال بهینه‌سازی مبتنی‌بر روش Limited memory-BFGS، نتایج همدوسی کمتری را نسبت به سایر روش‌ها گزارش می‌دهد. در ادامه همدوسی متقابل میان اتم‌های واژه‌نامه گفتار و هر یک از واژه‌نامه‌های نوفه به‌کارگرفته‌شده در روال بهسازی مورد بررسی قرار گرفت. این مقادیر همدوسی در جدول (۲) گزارش شده است. در سه روش نخست یادگیری واژه‌نامه در نظر گرفته‌شده در این جدول، مقادیر همدوسی متقابل اتم‌های گفتاری و هر یک از نوفه‌ها مقادیر بالایی دارد؛ زیرا در این روش‌ها روالی برای کاهش این پارامتر درنظر گرفته نشده است؛ اما در روش پیشنهادی در این مقاله با توجه به رابطه (۱۲)، بخش از تابع بهینه‌سازی مبتنی‌بر کاهش مقدار همدوسی متقابل اتم‌های واژه‌نامه نوفه و گفتار درنظر گرفته شده است. مقدار همدوسی متقابل واژه‌نامه گفتار و نوفه گزارش‌شده در این جدول، کارایی الگوریتم پیشنهادی را تأیید و منجر به نتایج بهسازی مناسبی می‌شود. در گام بهسازی نظارت‌شده، در ابتدا واژه‌نامه نوفه منطبق با شرایط محیط آزمون مطابق با آنچه در بخش ۳-۳ بیان شد، انتخاب می‌شود و در روال بهسازی مورد استفاده قرار می‌گیرد. روش‌های پایه بهسازی گفتار که مقایسه با آن‌ها صورت گرفته شامل روش تفاضل طیفی چندباند MBSS[1] [2]، تفاضل طیفی هندسی GSS[2] [41] و روش پایه آستانه‌گذاری موجک BWT[3] با تابع آستانه نرم معرفی‌شده در [3] است؛ همچنین عملکرد الگوریتم با روش ارائه‌شده در [43]، روش مبتنی‌بر یادگیری واژه‌نامه ارائه‌شده در [29] و نیز الگوریتم پیشنهادی در [44] مقایسه شده است.

نتایج حاصل از روش ارائه‌شده در [29] و روش پیشنهادی در [44] به‌ترتیب با نام GDL[4] و MDL[5] در شکل‌ها و جداول بیان می‌شوند؛ همچنین روش با نام AWPT[6] مربوط به الگوریتم ارائه‌شده در [43] است که بهسازی گفتار را براساس آستانه‌گذاری وفقی ضرایب تبدیل موجک انجام می‌دهد. علت ارائه این روش نواقصی بوده است که عملکرد الگوریتم آستانه‌گذاری ضرایب موجک پایه را محدود می‌کند؛ مانند: فرض نوفه گوسی سفید، عملکرد نامناسب در بخش‌های گفتاری، کیفیت شنیداری بد. در الگوریتم پیشنهادی، مقدار آستانه وفقی برای ضرایب موجک و تابع آستانه‌گذار اصلاح‌شده‌ای ارائه شده که دقت عملکرد بهسازی گفتار مناسبی دارد. همچنین آشکارساز فعالیت گفتاری جدیدی طراحی‌شده تا آمارگان نوفه را متناسب با سطح نوفه تخمین زده‌شده به‌روز کند. نتایج روش پیشنهادی در سناریوی نظارت‌شده با نماد اختصاری ((Proposed_Sup)) و در سناریوی نیمه‌نظارت‌شده با نماد ((Proposed_Semi)) بیان شده است. به‌منظور سهولت در بررسی نتایج گزارش‌شده، نام اختصاری روش‌های بیان‌شده که الگوریتم پیشنهادی با آن‌ها مقایسه می‌شود در جدول (۳) آورده شده است. معیارهای ارزیابی استفاده‌شده در این مقاله fwSegSNR[7] و PESQ[8] خواهند بود.

(جدول-۳): نام اختصاری روش‌های مقایسه‌شده با الگوریتم پیشنهادی به‌کاربرده‌شده در شکل‌ها و جداول.

(Table -3): The abbreviation name used for each comparison method in Figures and Tables.

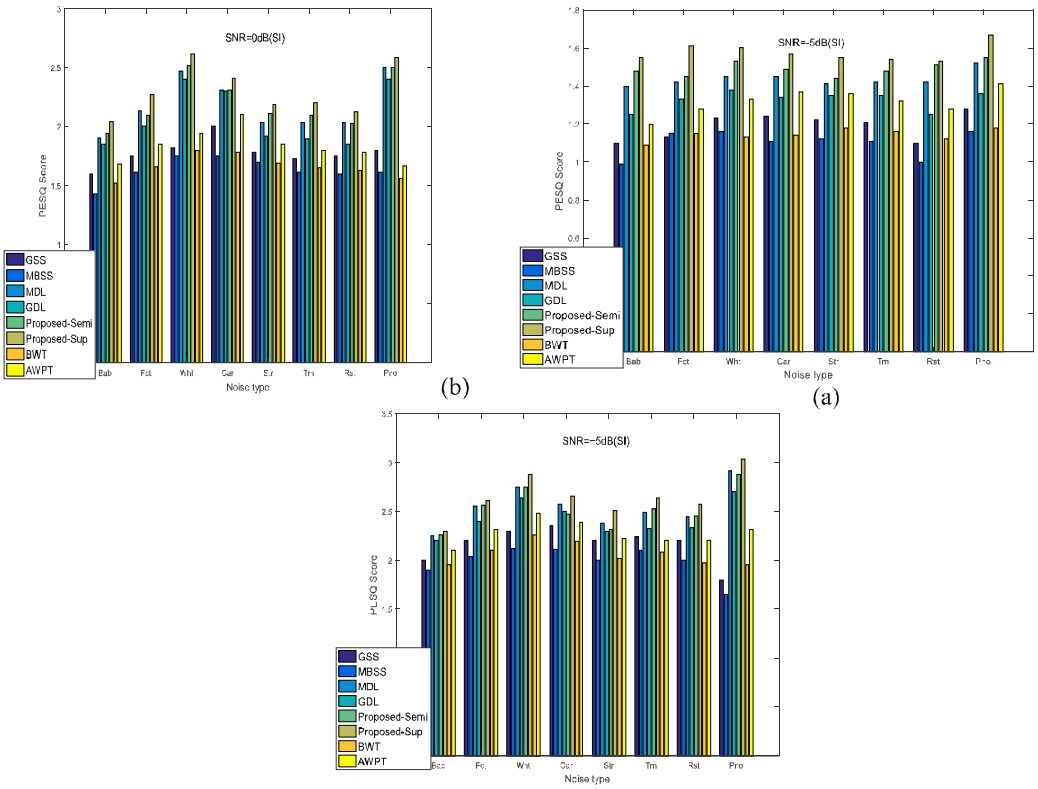
نام اختصاری	نام روش	مرجع
GSS	تفاضل طیفی هندسی	[40]
MBSS	تفاضل طیفی چندباند	[2]
MDL	یادگیری واژه‌نامه اصلاح شده	[42]
GDL	یادگیری واژه‌نامه مولد	[27]
BWT	روش پایه آستانه‌گذاری موجک	[3]
AWPT	بسته موجک تطبیقی	[41]
Proposed_Sup	روش پیشنهادی نظارت‌شده	-
Proposed_Semi	روش پیشنهادی نیمه نظارتی	-

[1] Multi band spectral subtraction

[2] Geometric spectral subtraction
[3] Basic wavelet thresholding algorithm(BWT)
[4] Generative dictionary learning(GDL)
[5] Modified dictionary learning(MDL)
[6] Adaptive wavelet packet thresholding
[7] Frequency-weighted segmental signal to noise ratio
[8] Perceptual evaluation of speech quality

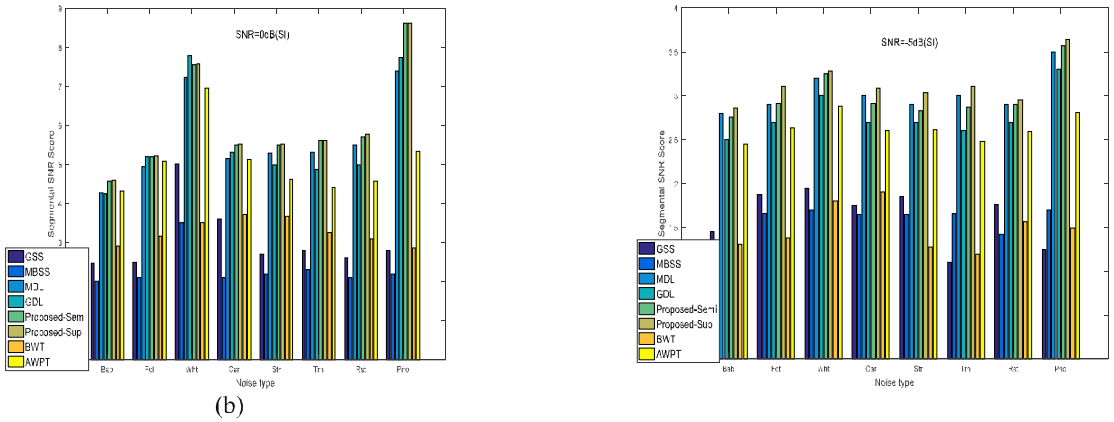
[DOI: 10.29252/jsdp.17.3.17]

نتایج بهسازی حاصل از روش‌های یادشده برای این دو معیار در حالت مستقل از گوینده(SI) در سه مقدار SNR، ۵dB-، ۰dB و ۵dB+ و برای هشت نوع نوفه: همهمه(Bab)، کارخانه(Fct)، سفید(Wht)، ماشین(Car)، خیابان(Str)، قطار(Trn)، رستوران(Rst) و نوفه موسیقی پیانو (Pno) در شکل‌های (۵ و ۶) نشان داده شده است. این نتایج در مقادیر SNR، ۵dB-، ۰dB و ۵dB+ به‌دست آمده است. همچنین به‌منظور نمایش بهتر تفاوت میان دسته روش‌های مختلف در شکل‌های (۵ و ۶)، این نتایج در جداول (۴ و ۵) گزارش شده است. نتایج حاصل از این معیارها در حالت وابسته به گوینده (SD) نیز در جداول (۶ و ۷) بیان شده است. همان‌طورکه در شکل‌ها و جداول مشاهده می‌شود، روش پیشنهادی نتایج مطلوبی را در ارزیابی با هر دو معیار PESQ و fwSegSNR نسبت به سایر الگوریتم‌های پایه و روش‌های پیشین مبتنی‌بر یادگیری واژه‌نامه در شرایط نوفه‌ای متفاوت به‌دست می‌دهد.

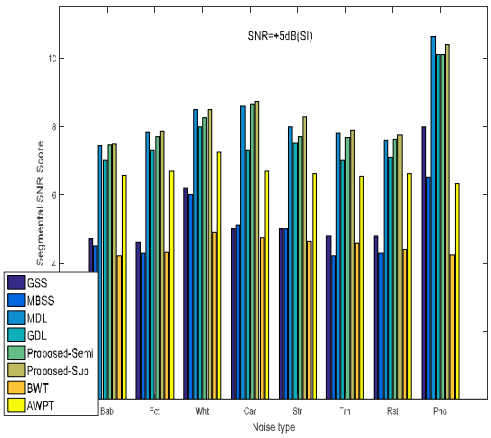


(شکل-۵): نتایج معیار PESQ به‌منظور مقایسه عملکرد روش‌های مختلف در سناریوSI و SNRهای: a) -5dB b) 0dB c) +5dB.

(Figure -5): Performance evaluation of different methods using PESQ score in speaker independent case (SI). a) SNR = -5dB, b) SNR = 0dB and c) SNR = +5dB.



[DOI: 10.29252/jsdp.17.3.17]



(c)

(شکل-۶): نتایج معیار SNR قطعه‌ای به‌منظور مقایسه عملکرد روش‌های مختلف در سناریوی SI و SNRهای: a) -5dB، b) 0dB و c) +5dB.

(Figure-6): Performance evaluation of different methods using frequency weighted segmental SNR value in speaker independent case (SI). a) SNR = -5dB, b) SNR = 0dB and c) SNR = +5dB.

همچنین میانگین نتایج بر روی تمامی شرایط نوفه‌ای (سیگنال‌های نوفه‌ای و SNRهای مختلف) برای این دو معیار و در شرایط وابسته و مستقل از گوینده به‌ترتیب در جداول (۸ و ۹) گزارش شده است.

نتایج مناسب حاصل از روش پیشنهادی برای تمامی نوفه‌های دارای ساختار و بدون ساختار و حتی در مقادیر SNR کم، قابل مشاهده است. همان‌طورکه انتظار می‌رفت، معیارهای بهسازی گزارش‌شده در حضور نوفه سفید و نوفه موسیقی بهترین مقادیر و در حضور نوفه‌های با ناایستایی بیشتر مانند نوفه همهمه و رستوران نتایج پایین‌تری را به‌دست می‌دهند. بر طبق انتظار و شبیه‌سازی‌های به‌دست‌آمده، نتایج حاصل از بهسازی در وضعیت وابسته به گوینده به‌علت هم‌پوشانی ایجادشده میان گوینده‌ها در محیط آموزش و آزمون، بهتر از نتایج به‌دست‌آمده در وضعیت مستقل از گوینده خواهد بود.

(جدول-۴): نتایج معیار PESQ به‌منظور مقایسه عملکرد روش‌های مختلف در سناریوی SI. نتایج روش پیشنهادی در وضعیت بانظارت و نیمه‌نظارت‌شده به‌ترتیب با نماد اختصاری ((Proposed_Sup)) و ((Proposed_Semi)) بیان شده است.

(Table-4): The results of PESQ measure values for performance evaluation of different methods in the presence of the mentioned noise signals in SI scenario. The proposed method is introduced using ((Proposed_Sup)) and ((Proposed_Semi)) in supervised and semi supervised situations, respectively.

	SNR=+۵								SNR=-۵							
	AWPT	BWT	Proposd_Sup	Proposed_Semi	GDL	MDL	MBSS	GSS	AWPT	BWT	Proposed_Sup	Proposed_Semi	GDL	MDL	MBSS	GSS
همهمه	۲/۱۰	۱/۹۵	**۲/۳۰**	۲/۲۶	۲/۱۱	۲/۲۰	۱/۹۰	۲	۱/۲۰	۱/۰۹	**۱/۵۵**	۱/۴۸	۱/۲۵	۱/۳۸	۰/۹۹	۱/۱۰
کارخانه	۲/۳۱	۲/۱۰	**۲/۶۱**	۲/۵۶	۲/۲۴	۲/۴۹	۲/۰۴	۲/۲۰	۱/۲۸	۱/۱۰	**۱/۶۱**	۱/۴۵	۱/۳۳	۱/۴۰	۱/۱۵	۱/۱۳
سفید	۲/۴۸	۲/۲۶	**۲/۸۸**	۲/۷۵	۲/۴۴	۲/۶۹	۲/۱۲	۲/۳۰	۱/۳۳	۱/۱۳	**۱/۶۰**	۱/۵۳	۱/۳۸	۱/۴۳	۱/۱۶	۱/۲۳
ماشین	۲/۳۹	۲/۱۹	**۲/۶۶**	۲/۴۷	۲/۳۸	۲/۴۵	۲/۱۱	۲/۳۵	۱/۳۷	۱/۱۴	**۱/۵۷**	۱/۴۹	۱/۳۴	۱/۴۳	۱/۱۴	۱/۲۴
خیابان	۲/۲۲	۲/۰۲	**۲/۵۱**	۲/۳۱	۲/۲۷	۲/۳۱	۲/۰۵	۲/۲۰	۱/۳۶	۱/۱۸	**۱/۵۵**	۱/۴۴	۱/۳۱	۱/۳۸	۱/۱۲	۱/۲۲
قطار	۲/۲۰	۲/۰۸	**۲/۶۴**	۲/۵۳	۲/۲۳	۲/۴۳	۲/۱۰	۲/۲۴	۱/۳۲	۱/۱۶	**۱/۵۴**	۱/۴۸	۱/۳۳	۱/۴۱	۱/۱۱	۱/۲۱
رستوران	۲/۱۸	۱/۹۷	**۲/۵۷**	۲/۴۵	۲/۲۲	۲/۲۹	۲	۲/۱۴	۱/۲۸	۱/۱۲	**۱/۵۳**	۱/۵۱	۱/۲۶	۱/۴۰	۱	۱/۱۰
پیانو	۲/۳۱	۱/۹۵	**۳/۰۳**	۲/۸۸	۲/۶۵	۲/۸۲	۱/۶۵	۱/۸۰	۱/۴۱	۱/۱۸	**۱/۶۷**	۱/۵۵	۱/۳۶	۱/۵۳	۱/۱۶	۱/۲۸

[DOI: 10.29252/jsdp.17.3.17]

(جدول-۵): نتایج معیار SNR قطعه‌ای به منظور مقایسه عملکرد روش‌های مختلف در سناریوی SI. نتایج روش پیشنهادی در وضعیت بانظارت و نیمه‌نظارت‌شده به‌ترتیب با نماد اختصاری ((Proposed_Sup)) و ((Proposed_Semi)) بیان شده است.

(Table-5): The results of frequency weighted segmental SNR values for performance evaluation of different methods in the presence of the mentioned noise signals in SI scenario. The proposed method is introduced using ((Proposed_Sup)) and ((Proposed_Semi)) in supervised and semi supervised situations, respectively.

	SNR=+۵								SNR=-۵							
	AWPT	BWT	Proposed_Sup	Proposed_Semi	GDL	MDL	MBSS	GSS	AWPT	BWT	Proposed_Sup	Proposed_Semi	GDL	MDL	MBSS	GSS
همهمه	۶/۵۶	۴/۲۰	**۷/۴۸**	۷/۴۶	۷	۷/۴۳	۴/۵۰	۴/۷۰	۲/۴۵	۱/۳۱	**۲/۸۶**	۲/۷۶	۲/۵۰	۲/۸۰	۱/۳۵	۱/۴۵
کارخانه	۶/۷۱	۴/۳۲	**۷/۸۶**	۷/۶۹	۷/۳۰	۷/۸۲	۴/۳۲	۴/۶۱	۲/۶۳	۱/۳۸	**۳/۱۱**	۲/۹۱	۲/۷۰	۲/۹۰	۱/۶۶	۱/۸۷
سفید	۷/۲۵	۴/۸۹	**۸/۵۰**	۸/۲۵	۸/۰۳	۸/۴۳	۶/۰۲	۶/۲۱	۲/۸۸	۱/۸۰	**۳/۲۸**	۳/۲۵	۳	۳/۲۰	۱/۷۰	۱/۹۵
ماشین	۶/۶۸	۴/۷۵	**۸/۷۴**	۸/۶۵	۷/۳۰	۸/۶۰	۵/۱۰	۵/۰۳	۲/۶۰	۱/۹۰	**۳/۰۹**	۲/۹۱	۲/۷۰	۳	۱/۶۵	۱/۷۵
خیابان	۶/۶۱	۴/۶۲	**۷/۲۸**	۷/۶۹	۷/۵۲	۸/۰۴	۵/۰۸	۵/۰۶	۲/۶۱	۱/۲۸	**۳/۰۳**	۲/۸۳	۲/۷۰	۲/۹۰	۱/۶۵	۱/۸۵
قطار	۶/۵۴	۴/۵۹	**۷/۸۹**	۷/۶۸	۷/۰۲	۷/۸۳	۴/۲۲	۴/۸۱	۲/۴۸	۱/۱۹	**۳/۱۱**	۲/۸۷	۲/۶۰	۳	۱/۶۶	۱/۱۰
رستوران	۶/۶۲	۴/۴۳	**۷/۷۵**	۷/۶۳	۷/۱۲	۷/۶۱	۴/۳۲	۴/۸۱	۲/۵۹	۱/۵۶	**۲/۹۵**	۲/۹۰	۲/۷۰	۲/۹۰	۱/۴۲	۱/۷۶
پیانو	۶/۳۳	۴/۲۳	۱۰/۴	۱۰/۱	۱۰/۱	**۱۰/۸**	۶/۵۲	۸/۰۶	۲/۸۱	۱/۴۹	**۳/۶۴**	۳/۵۷	۳/۳۰	۳/۵۰	۱/۷۰	۱/۲۵

(جدول-۶): نتایج معیار PESQ به‌منظور مقایسه عملکرد روش‌های مختلف در سناریوی SD. نتایج روش پیشنهادی در وضعیت بانظارت و نیمه‌نظارت‌شده به‌ترتیب با نماد اختصاری ((Proposed_Sup)) و ((Proposed_Semi)) بیان شده است.

(Table-6): The results of PESQ measure values for performance evaluation of different methods in the presence of the mentioned noise signals in SD scenario. The proposed method is introduced using ((Proposed_Sup)) and ((Proposed_Semi)) in supervised and semi supervised situations, respectively.

	SNR=+۵								SNR=-۵							
	AWPT	BWT	Proposed_Sup	Proposed_Semi	GDL	MDL	MBSS	GSS	AWPT	BWT	Proposed_Sup	Proposed_Semi	GDL	MDL	MBSS	GSS
همهمه	۲/۱۰	۱/۹۵	**۲/۳۷**	۲/۳۹	۲/۱۹	۲/۳۰	۱/۹۰	۲	۱/۲۰	۱/۰۹	**۱/۵۸**	۱/۵۱	۱/۳۳	۱/۴۶	۰/۹۹	۱/۱۰
کارخانه	۲/۳۱	۲/۱۰	**۲/۶۹**	۲/۵۹	۲/۳۱	۲/۵۶	۲/۰۴	۲/۲۰	۱/۲۸	۱/۱۰	**۱/۶۳**	۱/۵۳	۱/۴۲	۱/۴۹	۱/۱۵	۱/۱۳
سفید	۲/۴۸	۲/۲۶	**۲/۸۸**	۲/۸۰	۲/۴۹	۲/۷۵	۲/۱۲	۲/۳۰	۱/۳۳	۱/۱۳	**۱/۶۹**	۱/۶۰	۱/۴۷	۱/۵۴	۱/۱۶	۱/۲۳
ماشین	۲/۳۹	۲/۱۹	**۲/۷۲**	۲/۵۹	۲/۴۶	۲/۵۶	۲/۱۱	۲/۳۵	۱/۳۷	۱/۱۴	**۱/۶۴**	۱/۵۸	۱/۴۵	۱/۵۲	۱/۱۴	۱/۲۴
خیابان	۲/۲۲	۲/۰۲	**۲/۵۶**	۲/۵۰	۲/۳۵	۲/۴۵	۲/۰۵	۲/۲۰	۱/۳۶	۱/۱۸	**۱/۶۲**	۱/۵۵	۱/۴۰	۱/۴۶	۱/۱۲	۱/۲۲
قطار	۲/۲۰	۲/۰۸	**۲/۶۹**	۲/۵۸	۲/۴۰	۲/۵۲	۲/۱۰	۲/۲۴	۱/۳۲	۱/۱۶	**۱/۵۹**	۱/۵۴	۱/۴۴	۱/۵۰	۱/۱۱	۱/۲۱
رستوران	۲/۱۸	۱/۹۷	**۲/۶۴**	۲/۴۸	۲/۲۷	۲/۳۵	۲	۲/۱۴	۱/۲۸	۱/۱۲	**۱/۵۷**	۱/۵۳	۱/۳۷	۱/۴۸	۱	۱/۱۰
پیانو	۲/۳۱	۱/۹۵	**۳/۰۲**	۲/۸۹	۲/۶۸	۲/۸۶	۱/۶۵	۱/۸۰	۱/۴۱	۱/۱۸	**۱/۶۸**	۱/۵۸	۱/۴۵	۱/۵۶	۱/۱۶	۱/۲۸

(جدول-۷): نتایج معیار SNR قطعه‌ای به‌منظور مقایسه عملکرد روش‌های مختلف در سناریوی SD. نتایج روش پیشنهادی در وضعیت بانظارت و نیمه‌نظارت‌شده به‌ترتیب با نماد اختصاری ((Proposed_Sup)) و ((Proposed_Semi)) بیان شده است.

(Table-7): The results of frequency weighted segmental SNR values for performance evaluation of different methods in the presence of the mentioned noise signals in SD scenario. The proposed method is introduced using ((Proposed_Sup)) and ((Proposed_Semi)) in supervised and semi supervised situations, respectively.

	SNR=+۵								SNR=-۵							
	AWPT	BWT	Proposed_Sup	Proposed_Semi	GDL	MDL	MBSS	GSS	AWPT	BWT	Proposed_Sup	Proposed_Semi	GDL	MDL	MBSS	GSS
همهمه	۶/۵۶	۴/۲۰	**۷/۶۶**	۷/۵۴	۷/۲۳	۷/۵۸	۴/۵۰	۴/۷۰	۲/۴۵	۱/۳۱	**۳/۱۸**	۳/۰۹	۲/۵۷	۲/۹۴	۱/۳۵	۱/۴۵
کارخانه	۸/۷۱	۴/۳۲	**۸/۰۲**	۷/۹۹	۷/۴۵	۷/۹۱	۴/۳۲	۴/۶۱	۲/۶۳	۱/۳۸	**۳/۲۰**	۳/۱۴	۲/۷۰	۳/۱۱	۱/۶۶	۱/۸۷

[DOI: 10.29252/jsdp.17.3.17]

سفید	۷/۲۵	۴/۸۹	**۸/۷۹**	۸/۶۶	۸/۲۲	۸/۶۲	۶/۰۲	۶/۲۱	۲/۸۸	۱/۸۰	**۳/۳۷**	۳/۲۸	۳	۳/۴۱	۱/۷۰	۱/۹۵
ماشین	۶/۶۸	۴/۷۵	**۸/۷۶**	۸/۷۱	۷/۴۱	۸/۶۹	۵/۱۰	۵/۰۳	۲/۶۰	۱/۹۰	**۳/۲۹**	۳/۲۲	۲/۹۲	۳/۱۷	۱/۶۵	۱/۷۵
خیابان	۶/۶۱	۴/۶۲	**۸/۴۸**	۸/۳۶	۷/۶۲	۸/۳۱	۵/۰۸	۵/۰۶	۱/۶۲	۱/۲۸	**۳/۲۴**	۳/۱۸	۲/۸۸	۳/۱۶	۱/۶۵	۱/۸۵
قطار	۶/۵۴	۴/۵۹	**۸/۰۱**	۷/۹۱	۷/۲۳	۷/۸۹	۴/۲۲	۴/۸۱	۲/۴۸	۱/۱۹	**۳/۲۳**	۳/۰۹	۲/۷۵	۳/۱۱	۱/۶۶	۱/۱۰
رستوران	۶/۶۲	۴/۴۳	**۷/۸۶**	۷/۸۰	۷/۳۵	۷/۷۵	۴/۳۲	۴/۸۱	۲/۵۹	۱/۵۶	**۳/۱۵**	۳/۰۸	۲/۶۱	۳/۰۷	۱/۴۲	۱/۷۶
پیانو	۶/۳۳	۴/۲۳	۱۰/۶	۱۰/۵	۱۰/۳	**۱۰/۹**	۶/۵۲	۸/۰۶	۲/۸۱	۱/۴۹	**۳/۷۹**	۳/۶۶	۳/۵۱	۳/۶۴	۱/۷۰	۱/۲۵

(جدول-۸): میانگین نتایج معیار PESQ در شرایط مختلف نوفه‌ای به‌منظور مقایسه عملکرد روش پیشنهادی در سناریوی SI و SD.

(Table-8): The average results of PESQ score of the proposed method in SI and SD scenarios in the presence of the mentioned noise signals.

		GSS	MBSS	MDL	GDL	Proposed_Semi	Proposed_Sup	BWT	AWPT
PESQ	SD	۱/۷۹	۱/۶۰	۲/۰۹	۱/۹۶	۱/۸۶	۱/۹۵	۱/۵۵	۱/۶۷
	SI	۱/۷۹	۱/۶۰	۲/۰۱	۱/۸۸	۱/۷۹	۱/۸۹	۱/۵۵	۱/۶۷

(جدول-۹): میانگین نتایج معیار SNR قطعه‌ای در شرایط مختلف نوفه‌ای به‌منظور مقایسه عملکرد روش پیشنهادی در سناریوی SI و SD.

(Table-9): The average results of frequency weighted segmental SNR values of the proposed method in SI and SD scenarios in the presence of the mentioned noise signals.

		GSS	MBSS	MDL	GDL	Proposed_Semi	Proposed_Sup	BWT	AWPT
PESQ	SD	۳/۵۱	۳/۳۰	۴/۴۴	۴/۳۱	۴/۵۳	۴/۶۵	۳/۳۰	۴/۰۶
	SI	۳/۵۱	۳/۳۰	۴/۳۰	۴/۱۶	۴/۳۵	۴/۴۷	۳/۳۰	۴/۰۶

(جدول-۱۰): نتایج معیار PESQ و SNR قطعه‌ای در سناریوی SI به‌منظور مقایسه عملکرد روش پیشنهادی در وضعیت نیمه‌نظارت‌شده، روش پیشنهادی نیمه‌نظارت‌شده بدون گام آشکارساز فعالیت گفتاری VAD و نیز روش پیشنهادی نیمه‌نظارت‌شده بدون گام تطبیق‌فضا.

(Table-10): The results of PESQ and frequency weighted segmental SNR measures in SI scenario for performance evaluation of the proposed semi supervised method, the proposed semi supervised method without VAD step, and the proposed semi supervised method without domain adaptation technique.

			همهمه	کارخانه	سفید	ماشین	خیابان	قطار	رستوران	پیانو
Proposed_Semi	PESQ	+۵	۲/۲۶	۲/۵۶	۲/۷۵	۲/۴۷	۲/۳۱	۲/۵۳	۲/۴۵	۲/۸۸
		−۵	۱/۴۸	۱/۴۵	۱/۵۳	۱/۴۹	۱/۴۴	۱/۴۸	۱/۵۱	۱/۵۵
	SegSNR	+۵	۷/۴۶	۷/۶۹	۸/۲۵	۸/۶۵	۷/۶۹	۷/۶۸	۷/۶۳	۱۰/۱
		−۵	۲/۷۶	۲/۹۱	۳/۲۵	۲/۹۱	۲/۸۳	۲/۸۷	۲/۹۰	۳/۵۷
Proposed_Semi بدون VAD	PESQ	+۵	۱/۹۱	۲/۰۸	۲/۶۰	۲/۱۳	۲/۰۷	۲/۱۲	۲/۱۶	۲/۳۹
		−۵	۱/۱۸	۱/۱۳	۱/۳۷	۱/۱۷	۱/۱۵	۱/۱۸	۱/۲۱	۱/۳۹
	SegSNR	+۵	۶/۹۹	۷/۲۸	۷/۹۸	۷/۳۱	۷/۳۳	۷/۲۶	۷/۳۱	۹/۸۶
		−۵	۲/۳۶	۲/۶۸	۳/۱۴	۲/۶۲	۲/۵۱	۲/۴۴	۲/۵۱	۳/۴۱
Proposed_Semi بدون گام تطبیق‌فضا	PESQ	+۵	۲/۰۴	۲/۲۷	۲/۶۶	۲/۲۳	۲/۱۴	۲/۲۵	۲/۲۸	۲/۷۱
		−۵	۱/۲۶	۱/۲۲	۱/۴۸	۱/۲۴	۱/۲۲	۱/۲۳	۱/۳۳	۱/۴۸
	SegSNR	+۵	۷/۱۸	۷/۳۵	۸/۰۹	۷/۴۵	۷/۴۳	۷/۴۴	۷/۴۸	۹/۹۷
		−۵	۲/۵۰	۲/۷۷	۳/۲۰	۲/۷۴	۲/۶۱	۲/۵۶	۲/۶۳	۳/۴۹

به‌منظور بررسی عملکرد بخش‌های مختلف الگوریتم پیشنهادی به‌صورت مجزا، آزمایشی ترتیب داده شده و در آن کارایی بلوک آشکارساز فعالیت گفتاری پیشنهادی و نیز گام تطبیق فضای به‌کارگرفته‌شده مورد ارزیابی قرار گرفته است. در این شبیه‌سازی، نتایج معیار PESQ و SNR قطعه‌ای در سناریوی SI به‌منظور مقایسه عملکرد روش پیشنهادی در وضعیت نیمه‌نظارت‌شده، روش پیشنهادی نیمه‌نظارت‌شده بدون گام آشکارساز فعالیت گفتاری VAD و نیز روش پیشنهادی نیمه‌نظارت‌شده بدون گام تطبیق‌فضا در جدول (۱۰) بیان شده است. همان‌طورکه از این نتایج مشخص است، این دو گام و به‌خصوص استفاده از گام آشکارسازی فعالیت گفتاری پیشنهادی به‌طور محسوسی در دست‌یابی به نتایج مطلوب مؤثر خواهد بود. همان‌طورکه از نتایج حاصل مشاهده می‌شود، تأثیر روش تطبیق فضا در دستیابی به نتایج حاصل نیز محسوس است و استفاده از این روش مناسب وضعیتی خواهد بود که یک واژه‌نامه نوفه اولیه در دسترس باشد. همچنین می‌توان مشاهده کرد که تأثیر این دو گام در حضور نوفه دارای ایستایی بیشتر مانند نوفه

[DOI: 10.29252/jsdp.17.3.17]

سفید، پریودیک و پیانو بسیار کمتر از سایر نوفه‌های دارای ناایستایی بیشتر است. این مسأله بدین دلیل است که نوفه‌های سفید و پیانو به‌علت دارابودن ایستایی مناسب خوش‌ساختار بوده و با همان واژه‌نامه آموزش‌دیده اولیه قابلیت بازنمایی مناسب را دارا هستند و دیگر نیاز به به‌روزرسانی اتم‌ها بر طبق شرایط محیط آزمون وجود ندارد زیرا شرایط محیط آموزش و نویز در مورد این دسته نوفه‌ها دارای شباهت بیشتری خواهد بود.

نتایج حاصل از الگوریتم پیشنهادی در این مقاله که مبتنی‌بر یادگیری واژه‌نامه مجزا برای هر یک از زیرباندهای تقریب و جزئیات است، در سناریوی نظارت‌شده بهتر از نتایج سایر روش‌ها بوده است. همچنین نتایج حاصل از سناریوی نیمه‌نظارت‌شده نیز اندکی کمتر از سناریوی نظارت‌شده است. این برتری می‌تواند به‌دلیل بازنمایی هر قاب مشاهده در زیرباندهای مختلف باشد که امکان نمایش مناسب‌تر محتوای سیگنال را به‌دست می‌دهد. این نتایج برای انواع نوفه ساختاریافته، متناوب و بدون ساختار قابل تعمیم است؛ همچنین عامل دیگری که موجب نتایج بهتر این روش نسبت به سایر روش‌ها می‌شود، تخمین نوفه‌ای است که در گام بهسازی صورت می‌پذیرد. این تخمین هم در سناریوی نظارت‌شده و هم نیمه‌نظارت‌شده انجام می‌شود. در سناریوی نظارت‌شده پس از به‌روزرسانی اتم‌های واژه‌نامه نوفه براساس قاب‌های نوفه استخراج‌شده از الگوریتم آشکارساز فعالیت گفتاری[1] مبتنی‌بر انرژی ماتریس ضرایب تُنُک مرحله قبل، شرایط نوفه‌ای فضای آموزش به فضای آزمون نزدیک‌تر شده و اتم‌های نوفه بازنمایی درستی از شرایط محیط در گام انتهایی بهسازی در این سناریو را به‌دست می‌دهند. در سناریوی نیمه‌نظارت‌شده نیز با توجه به تخمین قاب به قاب نوفه با توجه به رابطه (۱۷) و تعیین مقدار آستانه و تابع آستانه‌گذار براساس سطح واریانس نوفه فضای آزمون، حذف نوفه با توجه به شرایط نوفه‌ای محیط آزمون انجام می‌گیرد. در روش MDL که مبتنی بر یادگیری واژه‌نامه است، از واژه‌نامه‌های نوفه یادگیری‌شده در گام آموزش به‌صورت مستقیم در گام بهسازی استفاده می‌شود. در این روش به‌علت عدم انطباق شرایط نوفه‌ای در محیط آموزش و آزمون به‌خصوص در حضور نوفه‌های ناایستا، نتایج حتی در سناریوی نیمه‌نظارت‌شده، پایین‌تر از روش ارائه‌شده در این مقاله خواهد بود؛ همچنین روش پیشنهادی نسبت به روش بهسازی گفتار در فضای تبدیل موجک با تخمین نوفه به‌روزشونده AWPT، به‌علت استفاده از مدل‌های ناهمدوس برای مؤلفه‌های داده ورودی نتایج بهتری را به دست می‌دهد.

در ادامه به‌منظور تصمیم‌گیری دقیق در مورد کارایی روش‌های مختلف مورد بررسی در شرایط متنوع یادشده، آزمون معناداری آماری به‌کارگرفته شده است. در این مقاله از آزمون Friedmanبه همراه آزمون Holms post hoc برای مقایسه آماری نتایج بیش از دو الگوریتم استفاده شده است [45,46]. این آزمون بر روی هر هشت روش مورد بررسی در شکل‌ها و جداول بخش شبیه‌سازی و در حضور هشت نوع سیگنال نوفه، سه مقدار SNR (۵dB-، ۰dB و ۵dB+) و دو مورد وضعیت SI و SD انجام می‌شود؛ بنابراین در این آزمون، تعداد روش‌ها و شرایط مختلف به‌ترتیب J=۸ و I =۴۸ است. گفتنی است که هر چه تعداد شرایط در مقایسه با تعداد روش‌ها بیشتر باشد، نتایج مناسب‌تری حاصل می‌شود. آزمون Friedman غیرپارامتری یکی از بهترین روش‌ها به منظور مقایسه چند روش در مجموعه شرایط گوناگون است و به فرضیات اولیه نیاز ندارد. به‌عنوان مرور کلی بر روال این آزمون فرض می‌کنیم که R_j میانگین رتبه عملکرد برای j-اُمین روش از میان J روش مورد بررسی باشد که هر یک از معیارهای عینی یا ذهنی به‌دست آمده و این نتایج در I شرایط مختلف صورت پذیرفته باشد:

$$R_j = \frac{1}{I}\sum_{i=1}^{I} r_{ij} \qquad (۲۲)$$

که در آن r_{ij} رتبه عملکرد j-اُمین روش در وضعیت i-ام آزمون خواهد بود. الگوریتم با بهترین عملکرد در این آزمون، کمترین مقدار رتبه را خواهد داشت. آزمون معنادار آماری با این فرضیه پوچ[2] که همه الگوریتم‌ها کارایی یکسان دارند شروع به کار می‌کند [47]. با استفاده از آزمون آماری Friedman، می‌توان تصمیم گرفت که این فرضیه رد یا پذیرفته می‌شود. این آزمون به‌صورت زیر بررسی می‌شود:

$$\chi_F^2 = \frac{12I}{J(J+1)}\left[\sum_{j=1}^{J} R_j^2 - \frac{J(J+1)^2}{4}\right] \qquad (۲۳)$$

همچنین آمار اصلاح‌شده از آزمون Friedman که براساس توزیع F است با (J-۱) و (J-۱)×(I-۱) درجه آزادی تعریف می‌شود [47]:

[1] Voice activity detector (VAD)

[2] Null-hypothesis

[DOI: 10.29252/jsdp.17.3.17]

$$F_F = (I-1)\chi_F^2/(I(J-1)-\chi_F^2) \quad (۲۴)$$

در این آزمون آماری، اگر مقدار F_F، بزرگ‌تر از مقدار بحرانی χ_F^2 باشد، فرضیه پوچ رد خواهد شد؛ یعنی کارایی این J الگوریتم مشابه نیست؛ بنابراین باید تعیین شود که کدام الگوریتم عملکرد بهتری دارد که این کار توسط آزمون post-hoc انجام می‌گیرد. پارامتر Z_j برای هر روش مورد ارزیابی این آزمون به‌صورت زیر محاسبه می‌شود [47]:

$$Z_j = (R_0 - R_j)/\sqrt{J(J+1)/6I} \quad (۲۵)$$

R_0 مربوط به روشی است که بیشترین رتبه میانگین(پایین‌ترین سطح عملکرد) را دارد. مقدار Z به منظور محاسبه مقدار ρ در سطح معناداری آماری α=۰/۰۵ مورد استفاده قرار می‌گیرد. در چگالی احتمال توزیع نرمال استاندارد، مقدار ρ متناظر با سطح زیر این توزیع و خارج از محدوده $(-Z, Z)$ به‌صورت زیر خواهد بود:

$$\rho = 1 - \int_{-Z}^{Z} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \quad (۲۶)$$

نتایج شبیه‌سازی با استفاده از آزمون آماری براساس معیار ارزیابی Holm برای روش‌های بهسازی گفتار معرفی‌شده در تمامی شرایط نوفه‌ای با استفاده از معیارهای PESQ و SNR قطعه‌ای، در جدول (۱۱و ۱۲) بیان شده است. مقادیر پارامتر χ_F^2 حاصل از این جداول براساس رابطه (۲۳) به‌ترتیب برابر با ۴۵/۸۱، ۵۸/۳۷ است؛ همچنین مقدار پارامتر F_F با توجه به رابطه (۲۴) در این جداول به‌ترتیب ۲۳/۶۲، ۲۲/۰۸ خواهد بود. مقدار F بحرانی با درجه‌های آزادی (۱-۸) و (۱-۴۸)×(۱-۸) برابر با ۲/۰۳۷۴ است. از مقایسه این نتایج مشخص می‌شود که تمامی مقادیر F_F به‌دست آمده در هر آزمون آماری گزارش‌شده در این جداول، بیشتر از مقدار F بحرانی است؛ بنابراین بر طبق آزمون Friedman، فرضیه اولیه یکسان‌بودن کارایی آماری روش‌های مدنظر نادرست بوده و می‌توان ادامه آزمون post-hocرا به‌منظور تعیین رتبه کارایی روش‌ها دنبال کرد. همان‌طورکه از پیش بیان شد، R_0 مربوط به روشی است که بیشترین رتبه میانگین (پایین‌ترین سطح عملکرد) را در میان روش‌های مختلف دارد که در بررسی‌های انجام‌شده متعلق به روش BWTاست که به الگوریتم پایه بهسازی گفتار مبتنی‌بر تبدیل موجک می‌پردازد. به‌منظور محاسبه پارامتر Z برای محاسبه مقدار ρ در سطح معناداری آماری α=۰/۰۵، از رابطه (۲۵) استفاده شده است؛ همچنین مقدار i در (α/(J-i)) به‌منظور محاسبه پارامتر Holm، برابر با شماره ردیف روش مورد بررسی در جداول از ۱ (روش با عملکرد بهتر) تا J-۱ (روش با عملکرد پایین‌تر) برای روش‌های مختلف بدون درنظرگرفتن روش متناظر با R_0 خواهد بود. در این جداول با توجه به اینکه مقدار ρ برای تمامی روش‌ها، کمتر از مقدار بحرانی Holm متناظر به‌دست آمده است، می‌توان نتیجه گرفت که روش‌های مورد بررسی به‌ترتیب قرار گرفته شده در جدول، از روش با بهترین عملکرد تا روش با بدترین عملکرد در زمینه بهسازی گفتار کارایی دارند. نتایج آزمون آماری بیان می‌کند که رتبه میانگین برای تمامی وضعیت‌های در نظر گرفته‌شده در شرایط آزمون برای روش پیشنهادی نظارت‌شده، نزدیک به یک بوده که موجب می‌شود این روش در جداول (۱۱ و ۱۲) در بالاترین ردیف قرار گیرد.

(جدول-۱۱): نتایج آزمون آماری برای معیار PESQ در شرایط و نوفه‌های مختلف.

(Table-11): The results of the statistical Friedman and Holm's post hoc test for PESQ score over different methods and conditions.

Methods	Average Rank (R_j)	Z	ρ-value	Holm(α/(J-i))
Proposed_Sup	۱/۲۸	۷/۹۱	۰	۰/۰۰۷۱
Proposed_Semi	۱/۹۱	۶/۴۸	۰	۰/۰۰۸۳
MDL	۲/۳۵	۵/۴۱	۰	۰/۰۱۰۰
GDL	۲/۹۶	۴/۳۳	۰	۰/۰۱۲۵
AWPT	۳/۲۱	۴/۲۸	۰	۰/۰۱۶۷
MBSS	۳/۶۷	۳/۲۲	۰/۰۰۱۳	۰/۰۲۵۰
GSS	۴/۱۲	۲/۱۸	۰/۰۲۹۳	۰/۰۵۰۰
BWT(R_0)	۲/۳۱	-	-	-

[DOI: 10.29252/jsdp.17.3.17]

(جدول-۱۲): نتایج آزمون آماری برای معیار SNR قطعه‌ای در شرایط و نوفه‌های مختلف.

(Table-12): The results of the statistical Friedman and Holm's post hoc test for of frequency weighted segmental SNR values over different methods and conditions.

Methods	Average Rank (R_j)	Z	ρ-value	Holm(α/(J-i))
Proposed_Sup	۱/۲۴	۸/۲۳	۰	۰/۰۰۷۱
Proposed_Semi	۲/۰۴	۶/۹۱	۰	۰/۰۰۸۳
MDL	۲/۵۶	۵/۹۲	۰	۰/۰۱۰۰
GDL	۲/۸۳	۵/۳۸	۰	۰/۰۱۲۵
AWPT	۳/۱۲	۴/۹۲	۰	۰/۰۱۶۷
MBSS	۳/۷۵	۳/۶۱	۰/۰۰۰۳	۰/۰۲۵۰
GSS	۱/۲۶	۲/۶۷	۰/۰۰۷۶	۰/۰۵۰۰
BWT(R_0)	۵/۵۲	-	-	-

بنابراین می‌توان نتیجه گرفت که روش پیشنهادی نظارت‌شده در فضای ویژگی تبدیل موجک برای بهسازی گفتار، بهترین رتبه میانگین در میان سایر روش‌ها را دارد و به‌صورت آماری بهتر از سایر الگوریتم‌ها عمل می‌کند. همچنین روش پیشنهادی نیمه نظارت‌شده در فضای ویژگی تبدیل موجک رتبه بعدی را به خود اختصاص می‌دهد. از آنجایی‌که در این جداول الگوریتم BWT با تابع آستانه‌گذاری نرم به‌عنوان روش با میانگین رتبه R_0 معرفی شده و مقدار پارامتر Z برای آن صفر به‌دست می‌آید، مقدار پارامتر ρ برای آن قابل محاسبه نخواهد بود.

۵- نتیجه‌گیری

در این مقاله، یادگیری واژه‌نامه به‌منظور دست‌یابی به دقت بالاتر در بهسازی سیگنال گفتار و در فضای تبدیل موجک بررسی شده است. به‌کمک این تبدیل، امکان تجزیه سیگنال در زیرباندهای مختلف تقریب و جزئیات سیگنال امکان‌پذیر خواهد بود. تجزیه سیگنال به زیرباندها می‌تواند اطلاعات دقیقی از محتوای سیگنال به‌دست دهد. یادگیری واژه‌نامه و یافتن اتم‌هایی که با تقریب مناسبی قادر به مدل‌کردن این زیرباندها هستند، می‌تواند تحلیل جزئی‌تری را به‌دست داده و نتایج مطلوبی را در زمینه بهسازی گفتار فراهم کند. همچنین دو سناریوی نظارت‌شده و نیمه‌نظارت‌شده ارائه شد که در هر یک، الگوریتم آشکارساز فعالیت گفتاری با توجه به شرط‌های مخصوص به خود بر پایه محاسبه انرژی ضرایب تُنُک حاصل از بازنمایی سیگنال مشاهده‌شده بر روی واژه‌نامه‌ها در گام آموزش، پیشنهاد شد. در سناریوی نیمه‌نظارت‌شده در ابتدا بازنمایی تُنُک قاب مشاهده‌شده بر روی واژه‌نامه گفتار انجام می‌شود و از روی ماتریس ضرایب تُنُک حاصل، قاب ورودی بازسازی می‌شود. هر چقدر خطای این تقریب، میان قاب مشاهده‌شده و قاب بازسازی‌شده، کمتر باشد به این معنا است که قاب ورودی بر روی واژه‌نامه گفتار بازنمایی تُنُک مناسب داشته و می‌توان نتیجه گرفت که قاب، ماهیت گفتاری خواهد داشت؛ سپس به‌منظور اطمینان بیشتر به تصمیم گرفته‌شده، انرژی بازنمایی تُنُک قاب ورودی بر روی قاب‌های ابتدایی و انتهایی سیگنال گفتار محاسبه می‌شود. کوچک‌بودن انرژی محاسبه‌شده، صحت نتیجه حاصل از شرط قبل را تأیید می‌کند. در سناریوی نظارت‌شده که واژه‌نامه نوفه نیز در دسترس است، پس از انتخاب واژه‌نامه مرتبط، انرژی ضرایب تُنُک حاصل از بازنمایی تُنُک در زیرباندهای مختلف محاسبه، سپس از بررسی‌های انجام‌شده در جهت تصمیم‌گیری درست در تعیین نوع قاب ورودی استفاده می‌شود. در ادامه نتایج خروجی آشکارساز بسته به سناریوی مدنظر، طی روال بهسازی پیشنهاده‌شده، سیگنال گفتار تخمینی را به‌دست خواهد داد. نتایج گزارش‌شده بر مبنای معیارهای مورد بررسی، توانایی این روش نسبت به سایر روش‌های پایه ارائه‌شده را در فضای ویژگی موجک و دیگر روش‌های بهسازی گفتار مبتنی‌بر واژه‌نامه نشان می‌دهد.

6- References ۶- مراجع

[1] M. Klein, P. Kabal, "*Signal subspace speech enhancement with perceptual post-filtering*", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process. (ICASSP)*, Vol. 1, pp. 537-540, 2002.

[2] S. Kamath, P. Loizou, "*A multi-band spectral subtraction method for enhancing speech corrupted by colored noise*", *In: Proc. IEEE Internat. Conf. Acoust. Speech Signal Process. (ICASSP*), Orlando, Florida, 2002.

[DOI: 10.29252/jsdp.17.3.17]

[DOI: 10.29252/jsdp.17.3.17]

[3] D.L. Donoho, "*De-noising by soft-thresholding*", *IEEE Trans. Inf. Theory*, Vol. 41, No. 3, pp. 613-627, 1995.

[4] N. Upadhyay, R.K. Jaiswal, "*Single Channel Speech Enhancement: Using Wiener Filtering with Recursive Noise Estimation*", *Procedia Computer Science*, Vol. 84, pp. 22-30, 2016.

[5] J. Candes, M.B. Wakin, "*An introduction to compressive sampling*", *IEEE Signal Processing Magazine*, pp. 21-30, 2008.

[6] R.G. Baraniuk, "*Compressive Sensing*", *IEEE Signal Processing Magazine*, pp. 118-121, 2007.

[7] S. Ayat, R. Dianat, M. Manzuri, "*Wavelet Based Speech Enhancement Using a New Thresholding Algorithm*", *IEEE Intl. Symposium on Intelligent Multimedia, Video & Speech Processing (ISIMP)*, Hong Kong, 2004.

[8] C.T. Lu, H.C. Wang, "Speech enhancement using wavelet transform with constrained thresholds", *In Proc. The 3rd International Symposium on Chinese Spoken Language Processing (ISCSLP)*, Taipei, Taiwan, pp. 185-188, 2002.

[9] E. Ambikairajah, G. Tattersall, A. Davis, "*Wavelet Transform-based Speech Enhancement*", *Proc. on ICSLP*, Vol. 3, 1998.

[10] V.S.R. Kumari, D.K. Devarakonda, "*A Wavelet Based Denoising of Speech Signal*", *International Journal of Engineering Trends and Technology (IJETT)*, Vol. 5, No. 2, pp. 107-115, 2013.

[11] K. Khaldi, A.O. Boudraa, A. Komaty, "Speech enhancement using empirical mode decomposition and the Teager-Kaiser energy operator", *J Acoust Soc Am*, Vol. 13, No. 5, pp. 451-459, 2014.

[12] T.F. Sanam, C. Shahnaz, "A semisoft thresholding method based on Teager energy operation on wavelet packet coefficients for enhancing noisy speech", *EURASIP Journal on Audio, Speech and Music Processing*, Springer, 2013.

[13] S. Hongo, S. Sakamoto, Y. Suzuki, "Binaural speech enhancement method by wavelet transform based on interaural level and argument differences", *International Conference on Wavelet Analysis and Pattern Recognition*, Xian, 2012, pp. 290-295.

[14] T. V. Pham, "Wavelet analysis for robust speech processing and applications", PHD Thesis, Graz University of Technology, 2007.

[15] I. Pinter, "Perceptual wavelet-representation of speech signals and its application to speech enhancement", *Computer Speech and Language*, Vol. 10, No. 1, pp. 1-22, 1996.

[16] M.A. Messaoud, A. Bouzid, "Speech enhancement based on wavelet transform and improved subspace decomposition", *Journal of Audio Engineering society (JAES)*, Vol. 63, No.12, pp.1-11, 2015.

[17] C.L. Wu, H.P. Hsu, S.S. Wang, J.W. Hung, Y.H. Lai, H.M. Wang, Y. Tsao, "Wavelet speech enhancement based on robust principal component analysis", *Proc. Interspeech* , 781, pp. 439-443, 2017.

[18] T.Y. Zuo, L. He, W.D. Sheng, "A new algorithm of the wavelet packet speech denoising based on masking perception model", *7th International conference on natural computation (ICNC)*, Vol. 1, 2011, pp. 33-37.

[19] H. Zhao, X. Peng, L. Hu, G. Wang, "An improved speech enhancement method based on teager energy operator and perceptual wavelet packet decomposition", *Journal of Multimedia*, Vol. 6, No. 3, 2011.

[20] R. Gomez, T. Kawahara, "Optimized wavelet-based speech enhancement for speech recognition in noisy and reverberant conditions", APSIPA ASC, 2011.

[21] T.F. Sanam, C. Shahnaz, "Teager energy operation on wavelet packet coefficients for enhancing noisy speech using a hard thresholding function", *Published in Signal Processing: An International Journal (SPIJ)*, Vol. 6, pp. 22-43, 2011.

[22] G. Chen, C. Xiong, J.J Corso, "Dictionary transfer for image denoising via domain adaptation", *In Proceedings of IEEE International Conference on Image Processing*, 2012.

[23] S. Mavaddati, S.M. Ahadi Sarkani, S. Seyedin, "A novel speech enhancement method by learnable sparse and low-rank decomposition and domain adaptation", *Speech Communication*, Vol. 76, pp. 42-60, 2016.

[24] A. Agarwal, A. Anandkumar, P. Jain, P. Netrapalli, R. Tandon, *JMLR: Workshop and Conference Proceedings*, Vol. 35, 2014, pp. 1-15.

[25] H. Lee, A. Battle, R. Raina, A.Y. Ng, "Efficient sparse coding algorithms", *Advances in Neural Information Processing Systems*, 2006.

[26] J. Portilla, L. Mancera, "L_0-based sparse approximation: Two alternative methods and some applications", *Proceedings of the 16th IEEE international conference on Image processing*, 2009, pp. 3865-3868.

[۲۷] مودّتی، سمیرا، احدی، محمد،"بهسازی گفتار به‌کمک یادگیری واژه‌نامه مبتنی‌بر داده"، پردازش علائم و داده‌ها، جلد ۱۷، شماره ۱، صفحه ۹۹-۱۱۶، ۱۳۹۹.

[27] S.Mavaddati, M. Ahadi, “Speech Enhancement using Adaptive Data-Based Dictionary Learning”, *JSDP*, vol. 17 (1), pp. 99-116, 2020.

[۲۸] مظفری، رضا، مودّتی، سمیرا، "ارائه روش جدید حذف نوفه تصویر براساس یادگیری واژه‌نامه ناهمدوس و روش تطبیق فضا"، پردازش علائم و داده‌ها، جلد ۱۶، شماره ۴، صفحه ۷۳-۹۲، ۱۳۹۸.

[28] R. Mozaffari, S. Mavaddati, “A Novel Image Denoising Method Based on Incoherent Dictionary Learning and Domain Adaptation Technique”, *JSDP*, vol. 16 (4), pp.73-92. 2020.

[29] C.D. Sigg, T. Dikk, J.M. Buhmann, "Speech enhancement using generative dictionary learning", *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 20, No.6, pp.1698-1712, 2012.

[30] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, "Least angle regression*", Ann. Stat.*, Vol. 32, pp. 407-499, 2004.

[31] M. Aharon, M. Elad, A. Bruckstein, "*K-SVD:* An algorithm for designing overcomplete dictionaries for sparse representation”, *IEEE Trans. Signal Process*, Vol. 54, No. 11, pp. 4311-4322, 2006.

[32] X. Wu, D. Yu, "Atomic Decomposition Method Based on Adaptive chirplet Dictionary", *Advances in Adaptive Data Analysis*, Vol. 4, pp. 1-19, 2012.

[33] J. Tropp, I. Dhillon, R.J. Heath, T. Strohmer, "Designing structural tight frames via an alternating projection method", *IEEE Trans. on Information Theory*, Vol. 51, No.1, pp. 188-209, 2005.

[34] M. Sustik, J. Tropp, I. Dhillon, R. Heath, "On the existence of equiangular tight frames", *Linear Algebra and Its Applications*, Vol. 426, No. 2, pp. 619-635, 2007.

[35] D. Liu, J. Nocedal, "On the limited memory BFGS method for large scale optimization", *Math. Program,* Vol. 45, pp. 503-528, 1989.

[36] S. Mavaddati, S.M. Ahadi Sarkani, S. Seydin, "Speech enhancement using sparse dictionary learning in wavelet packet transform domain", *Computer Speech and Language,* Vol. 44, pp. 22-47, 2017.

[37] D.L. Donoho, "De-noising by soft-thresholding", *IEEE Trans. Inf. Theory,* Vol. 4, No. 3, pp. 613-627, 1995.

[38] http://www.dcs.shef.ac.uk/spandh/gridcorpus.

[39] A. Varga, H. Steeneken, J.M. Tomlinson, D. Jones, "The Noisex-92 study on the effect of additive noise on automatic speech recognition", *Technical Report. Malvern*, U.K.: DRA Speech Res. Unit, 1992.

[40] H.G. Hirsch, D. Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions", *Proc. ISCA ITRWASR*, pp.181-188, 2000.

[41] http://pianosociety.com.

[42] Y. Lu, P.C. Loizou, "A geometric approach to spectral subtraction", *Speech communication*, Vol. 50, No. 6, pp. 453-466, 2008.

[43] Y. Ghanbari, M.R. Karami Mollaei, "A new approach for speech enhancement based on the adaptive thresholding of the wavelet packets", *Speech communication*, Vol. 48, No. 40, pp. 927-940, 2006.

[44] S. Mavaddati, S.M. Ahadi Sarkani, S. Seydin, "Modified coherence-based dictionary learning method for speech enhancement", *Signal Processing,* IET, Vol. 9, No. 7, pp. 1-9, 2015.

[45] J. Benesty, *Springer handbook of speech processing*, Springer’s publication, pp. 843-871, 2008.

[46] J. Demsar, "Statistical comparisons of classifiers over multiple data set", *The Journal of Machine Learning Research*, Vol. 7, pp. 1-30, 2006.

[47] D.J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, 4nd ed. Boca Raton, FL: Chapman & Hall/CRC, 2000.

سمیرا مودّتی مدارک کارشناسی و کارشناسی ارشد خود را به‌ترتیب در سال‌های ۱۳۸۶ و ۱۳۸۹ از دانشگاه مازندران در رشته مهندسی برق- الکترونیک و درجه دکترای خود را در سال ۱۳۹۵ از دانشگاه صنعتی امیرکبیر در رشته مهندسی برق- الکترونیک دریافت کرد. وی هم‌اکنون استادیار گروه مهندسی برق دانشگاه مازندران است. زمینه‌های پژوهشی مورد علاقه ایشان عبارت است از: پردازش سیگنال گفتار، پردازش سیگنال تصویر، بهینه‌سازی و هوش مصنوعی.

نشانی رایانامه ایشان عبارت است از:

s.mavaddati@umz.ac.i

[DOI: 10.29252/jsdp.17.3.17]