Semnan University

Journal of Modeling in Engineering

Journal homepage: https://modelling.semnan.ac.ir/

ISSN: 2783-2538

Research Article

Applying Dictionary Learning Algorithms In Sparse Representation of Speech Signals

Naser Sharafi [a], Salman Karimi [a,*] Samira Mavaddati [b]

[a] *Faculty of Engineering, Lorestan University, Khorramabad, Iran*
[b] *Faculty of Engineering and Technology, University of Mazandaran, Babolsar, Iran*

PAPER INFO

Paper history:

Received: 2024-05-07
Revised: 2025-03-29
Accepted: 2025-06-24

Keywords:

Sparse representation;
Dictionary learning;
Speech processing;
K-SVD;
OMP;
STFT.

ABSTRACT

As a widely used technique in signal processing, Sparse representation has gained significant attention in various fields, including data compression, noise reduction in speech and image signals, pattern recognition, and other signal processing-related issues. In such representations, signals are linearly combined using a small number of dictionary atoms, leading to data dimensionality reduction and improved signal processing efficiency. To accurately represent speech data, an appropriate dictionary is required to effectively represent speech signals' characteristics. In this paper, dictionaries are trained using dictionary learning algorithms and sparse representations such as MOD, K-SVD, RAMC, UD4-MOD, and OMP, in the time, time-frequency, and wavelet transform domains. The performance of the obtained dictionaries is evaluated using various time-frequency metrics such as RE, MSE, fwSegSNR, SegSNR, PESQ, and STOI. The results demonstrate that employing the K-SVD dictionary learning algorithm in conjunction with the OMP sparse representation algorithm in the STFT domain achieves promising results for speech signal reconstruction.

DOI: https://doi.org/10.22075/jme.2025.34065.2663

*** Corresponding author.**
E-mail address: karimi.salman@lu.ac.ir

How to cite this article:
N. Sharafi , S. Kkarimi and S. Mavaddati, "Applying Dictionary Learning Algorithms In Sparse Representation of Speech Data," Journal of Modeling in Engineering, 24 85 (2026): 17-32, doi: 10.22075/jme.2025.34065.2663

مقاله پژوهشی

بکارگیری الگوریتم‌های یادگیری واژه‌نامه در بازنمایی تُنُک دادگان گفتاری

ناصر شرفی[۱]، سلمان کریمی[۱*]، سمیرا مودّتی[۲]

اطلاعات مقاله

دریافت مقاله: ۱۴۰۳/۰۲/۱۸
بازنگری مقاله: ۱۴۰۴/۰۱/۰۹
پذیرش مقاله: ۱۴۰۴/۰۴/۰۳

واژگان کلیدی:
بازنمایی تُنُک،
آموزش واژه‌نامه،
پردازش سیگنال،
K-SVD،
OMP،
تبدیل فوریه کوتاه مدت.

چکیده

بازنمایی تُنُک به عنوان یکی از روش‌های پر کاربرد در پردازش سیگنال، در زمینه‌های مختلفی مانند فشرده‌سازی داده، حذف نویز از سیگنال‌های گفتاری و تصویری، تشخیص الگو و سایر مسائل مرتبط با پردازش سیگنال مورد توجه قرار گرفته است. در چنین بازنمایی‌هایی، سیگنال‌ها با استفاده از تعداد کمی از اتم‌های واژه‌نامه به‌صورت خطی ترکیب می‌شوند که منجر به کاهش ابعاد داده و بهبود کارایی در پردازش سیگنال می‌شود. به منظور بازنمایی دقیق‌تر داده‌های گفتاری، نیاز به واژه‌نامه مناسبی است که بتواند ویژگی‌های سیگنال گفتار را به خوبی نمایش دهد. در این مقاله، واژه‌نامه‌هایی با استفاده از الگوریتم‌های یادگیری واژه‌نامه و بازنمایی تُنُک MOD، K-SVD، RAMC و UD4-MOD و بازنمایی تُنُک OMP در حوزه‌های زمان، نمایش زمان-فرکانس و تبدیل موجک آموزش داده می‌شوند. ارزیابی کارایی واژه‌نامه‌های به‌دست‌آمده با استفاده از معیارهای مختلف زمانی و فرکانسی مانند RE، MSE، fwSegSNR، SegSNR، PESQ و STOI انجام شده است. نتایج حاصل، نشان می‌دهد که بکارگیری الگوریتم یادگیری واژه‌نامه K-SVD در ترکیب با الگوریتم بازنمایی تُنُک OMP در حوزه STFT نتایج مطلوبی را به منظور بازسازی سیگنال گفتاری به دست می‌دهد.

DOI: https://doi.org/10.22075/jme.2025.34065.2663

۱- مقدمه

یادگیری واژه‌نامه[۳] و بازنمایی تُنُک[۴]، دو روش مرتبط در یادگیری ماشین هستند که اغلب برای پردازش و تجزیه و تحلیل مجموعه داده‌های بزرگ کارایی دارند. یادگیری واژه‌نامه شامل یافتن مجموعه‌ای از توابع اتم یا پایه به نام واژه‌نامه است که می‌تواند تعداد زیادی نمونه آموزشی را بازنمایی نماید. واژه‌نامه آموخته شده مجموعه‌ای از بردارهایی است که ویژگی‌های کلیدی داده‌های آموزشی را به تصویر می‌کشد و می‌توان از آن برای نمایش داده‌های جدید و نادیده به صورت تُنُک و کارآمد استفاده نمود. بازنمایی تُنُک به روال نمایش یک بردار به عنوان ترکیبی از چند اتم واژه‌نامه اشاره دارد. با انتخاب تنها چند بردار از واژه‌نامه آموخته شده، می‌توان تعداد زیادی نمونه را در فضایی با ابعاد کم نمایش داد. این مفاهیم برای اولین بار در سال ۲۰۰۶ با عنوان حسگری فشرده[۵] معرفی گردید که کمک شایانی در توسعه نظریه بازنمایی تُنُک داشته است

* پست الکترونیک نویسنده مسئول: karimi.salman@lu.ac.ir
۱. دانشکده فنی و مهندسی، گروه مهندسی برق الکترونیک، دانشگاه لرستان، خرم‌آباد، ایران
۲. دانشکده مهندسی و فناوری، دانشگاه مازندران، بابلسر، ایران

[3] Dictionary Learning
[4] Sparse Representation
[5] Compressive Sensing

استناد به این مقاله:
ناصر شرفی , سلمان کریمی و سمیرا مودتی، "بکارگیری الگوریتم‌های یادگیری واژه‌نامه در بازنمایی تُنُک دادگان گفتاری،" مدل سازی در مهندسی، ۲۴ ۸۵ (۱۴۰۵): ۱۷-۳۲، doi: 10.22075/jme.2025.34065.2663

[۱]. کاربرد نمایش تُنُک به تدریج از فشرده‌سازی داده‌های اولیه به بسیاری از زمینه‌های پردازش تصویر و گفتار، پردازش زبان طبیعی و سیستم‌های توصیه‌گر گسترش یافت. برخی از این زمینه‌ها شامل حذف نویز، تشخیص الگو، تفسیر تصاویر پزشکی و پردازش گفتار می‌باشد که در آن‌ها داده‌ها به صورت ترکیبی از تعداد کمی از اجزاء پایه مدل می‌شوند[۲-۵]. در این مقاله کاربردهای این دو مفهوم به منظور بازنمایی سیگنال گفتار مورد بررسی قرار می‌گیرد. در این رابطه تاکنون، مطالعات متنوعی انجام شده است که در ادامه به بررسی آن‌ها پرداخته می‌شود. با استفاده از بازنمایی تُنُک و الگوریتم‌های یادگیری واژه‌نامه در [۶]، با هدف شناسایی نواحی غیر طبیعی مغز مبتلا به بیماری اسکیزوفرنی، واژه‌نامه جدیدی با ویژگی‌های مهم فشردگی گروهی[۶] و غیرهمدوسی[۷] میان اتم‌های آن طراحی شده است. نتایج بر روی تصاویر مغزی نشان‌دهنده دقت و حساسیت بالای این روش در شناسایی نواحی مهم عصبی می‌باشد. در [۷]، تصاویر با وضوح طیفی بالا (مانند تصاویر فراطیفی) با تصاویر با وضوح فضایی بالا ترکیب می‌شوند. این پژوهش براساس استراتژی ترکیب بیزین بوده که از اطلاعات پیشین و یادگیری واژه‌نامه برای بازنمایی نسبت داده به تصویر هدف که وضوح طیفی و فضایی بالایی دارد، استفاده می‌نماید. همچنین در این روش به توسعه روش‌های جدید تخمین حرکت بیزین برای تصاویر سونوگرافی پرداخته شده است. در [۸]، یک الگوریتم مبتنی‌بر یادگیری واژه‌نامه برای بهبود کیفیت گفتار از طریق بازنمایی تُنُک در حوزه تبدیل بسته موجک معرفی شده است. در این روش، یادگیری واژه‌نامه تُنُک برای داده‌های آموزشی گفتار و نویز براساس معیار همدوسی در هر زیرباند پیشنهاد شده است. این الگوریتم‌ها همدوسی درونی واژه‌نامه‌ها و همدوسی متقابل واژه‌نامه‌های گفتار و نویز را کاهش می‌دهد. الگوریتم در هر دو سناریوی نظارت‌شده و نیمه‌نظارت‌شده ارائه گردیده است [۸]. در [۹]، الگوریتم جدیدی به نام بازنمایی تُنُک در دامنه طیفی-زمانی[۸] (STRF) برای تشخیص فعالیت صوتی[۹] (VAD) ارائه شده است. این الگوریتم براساس مدل قشر شنوایی برای استخراج ویژگی و مدولاسیون طیفی-زمانی چند مقیاسی برای طبقه‌بندی صدا طراحی شده است. این الگوریتم گفتار را به حوزه طیفی-زمانی[۱۰] انتقال داده و از ویژگی‌های مبتنی‌بر شنوایی را با مقیاس‌های متعدد تفکیک زمانی و طیفی به دست می‌دهد. یک روش نوآورانه به منظور بهبود کیفیت گفتار با استفاده از نمایش تُنُک و مقیاس فرکانسی لگاریتمی در [۱۰] ارائه شده است. این روش با کاهش اندازه واژه‌نامه و بدون نیاز به تخمین نویز می‌تواند به طور مؤثری در شرایط نویز متغیر با نسبت‌های پایین سیگنال به نویز موثر بوده و کیفیت سیگنال گفتار را بهبود بخشد. با توجه به نتایج به دست آمده، این روش قابلیت اجرا در شرایط سخت محیطی را به اثبات رسانده و به عنوان یک پیشرفت قابل توجه در عرصه بهبود کیفیت گفتار مطرح می‌شود. یک الگوریتم جدید یادگیری واژه‌نامه مختص بهبود کیفیت گفتار در محیط‌هایی با نویز غیرایستا در [۱۱] معرفی شده است. این روش با تمرکز بر کاهش همدوسی بین زیرواژه‌نامه‌های گفتار و نویز و ارزیابی دقیق خطاهای مرتبط با اعوجاج منابع[۱۱]، در مرحله بهسازی از کدگذاری تُنُک استفاده می‌نماید تا تخمینی از طیف دامنه گفتار و نویز فراهم نماید. استفاده از فیلتر وینر در مرحله نهایی به اصلاح و بهبود بیشتر کیفیت گفتار کمک می‌نماید. در [۱۲]، یک روش جدید بهبود کیفیت گفتار مبتنی‌بر یادگیری واژه‌نامه و تجزیه ماتریس رتبه پایین[۱۲] پیشنهاد شده است. این روش با ارائه مدلی جدید از نویز و ترکیب آن با ماتریس کم‌رتبه و واژه‌نامه فراکامل[۱۳]، واژه‌نامه نویز را با همدوسی کمی نسبت به واژه‌نامه گفتار در مرحله یادگیری، آموزش می‌دهد. سپس مؤلفه کم‌رتبه نویز با الگوریتمی تحت‌نظارت، استخراج و خوشه‌بندی می‌شود. در نهایت در مرحله بهبود، گفتار تمیز از گفتار نویزی بازیابی می‌گردد. نتایج آزمایش‌ها حاکی از برتری این روش نسبت به روش‌های پایه در این حوزه پردازشی مانند الگوریتم‌های مبتنی‌بر یادگیری واژه‌نامه برای بهبود کیفیت گفتار است که در [۱۲] معرفی شده‌اند. سیستم VAD طراحی‌شده در هر سناریو براساس انرژی ضرایب تُنُک بکار گرفته شده است. هدف در [۱۳]، توسعه یک روش یادگیری واژه‌نامه فشرده است که خطای بازنمایی تُنُک را به حداقل می‌رساند و به چارچوب‌های سخت نُرم واحد[۱۴] (UNTF) نزدیک می‌شود. این امر با یک

[6] Group Sparsity
[7] Incoherence
[8] Sparse Representation in Spectro-Temporal Domain
[9] Voice Activity Detection
[10] Spectro-Temporal Domain

[11] Source Distortion
[12] Low Rank Matrix
[13] Overcomplete Dictionary
[14] Unit-Norm Tight Frame

رویکرد مبتنی‌بر گرادیان برای به دست آوردن واژه‌نامه فشرده که در آن محدودیت‌های عادی‌سازی در تابع هزینه با استفاده از یک مقداردهی اولیه احتمالی طراحی گردیده، حاصل شده است. واژه‌نامه فشرده آموخته‌شده به منظور بازیابی گفتار آسیب دیده بکار گرفته می‌شود تا سیگنال‌های گفتاری دارای اعوجاج توسط داده‌های از دست رفته را بازیابی نماید. در [۱۴]، روش جدیدی براساس تمایز بین آواهای صدادار و بی‌صدا در جملات پیشنهاد گردیده است. به منظور پیاده‌سازی این روش، نویسندگان الگوریتم تعقیب تطبیق متعامد[۱۵] (OMP) را برای دستیابی به نمایش پراکنده انتخاب نموده و سپس الگوریتم تجزیه K مقدار تکین[۱۶] (K-SVD) برای ساخت واژه‌نامه درنظر گرفته شده است. همانطور که بیان شد الگوریتم K-SVD به دلیل توانایی در مدیریت مجموعه داده‌های بزرگ و با ابعاد بالا می‌تواند در این راستا مفید باشد. در این روش، تعداد کافی از آواهای صدادار و بی‌صدا را برای یادگیری واژه‌نامه انتخاب و از K-SVD برای ساخت واژه‌نامه‌های مخصوص به هر دسته استفاده گردید. سپس، سیگنال‌های مورد ارزیابی به صورت نمایش تُنُک در هر دو واژه‌نامه بیان می‌گردند. شیوه تفکیک آواهای صدادار یا بی‌صدا در جمله، با مقایسه پراکندگی ضرایب در دو واژه‌نامه انجام و نتایج بر اثربخشی این روش تأیید می‌نماید. در مقاله [۱۵]، روش جدیدی با نام بازنمایی تُنُک خوشه‌بندی‌شده[۱۷] (CSSR) معرفی شده که به بهبود استقلال صدای گوینده در تبدیل صدا[۱۸] کمک می‌نماید. این روش با ترکیب دو جزء یادگیری واژه‌نامه ساختار یافته خوشه‌ای و تابع هدف انتخابی خوشه‌ای، کارایی واژه‌نامه و نمایش تُنُک را بهبود بخشیده و در نتیجه، عملکرد تبدیل صدا را افزایش می‌دهد. در [۱۶]، الگوریتم بهسازی گفتار مبتنی‌بر حسگری فشرده پیشنهاد شد که در آن نویز در مرحله اندازه‌گیری حذف شده و بازیابی به صورت فشرده انجام شده است. این روش از آموزش واژه‌نامه همراه با الگوریتم K-SVD برای ایجاد یک واژه‌نامه فراکامل استفاده می‌نماید که قادر به تخمین نویز در قاب‌های سکوت است. به منظور طبقه‌بندی قاب‌های گفتاری و سکوت، از الگوریتم VAD و یک تابع ماسک براساس خروجی این الگوریتم استفاده می‌شود. سپس، بردار مشاهده شده نویز تخمینی از بردار گفتار فعال کم می‌شود و با استفاده از بازیابی فشرده، طیف گفتار بهبود یافته به دست می‌آید. روشی جدید برای بهسازی سیگنال گفتاری به کمک روش‌های بازنمایی تُنُک و یادگیری واژه‌نامه در [۱۷] معرفی شده است. این روش از یک روال یادگیری ناهمدوس استفاده می‌نماید تا با استفاده از تعداد محدودی واژه‌نامه‌های یادگیری‌شده در فضای ویژگی‌های موجک، سیگنال نویز از سیگنال گفتاری حذف شود. با تجزیه سیگنال به زیرباندهای مختلف، دقیق‌ترین اطلاعات گفتاری به دست می‌آیند. در ادامه، روش‌های نظارت‌شده و نیمه‌نظارت‌شده برای بهسازی گفتار ارزیابی و الگوریتمی برای تشخیص VAD پایه‌گذاری شد تا در گام بعد منجر به بهسازی گفتار گردد. در [۱۸]، به منظور بهبود عملکرد سیستم‌های تشخیص گفتار[۱۹] (ASR)، سیگنال‌های گفتاری به بلوک‌های آکوستیکی مجزا تقسیم و با کمک تبدیل بسته موجک[۲۰] (WPT) به منظور پارامتری نمودن داده، گفتار برای تشخیص بهتر مرزهای واجی با استفاده از طبقه‌بند بازنمایی تُنُک[۲۱] (SRC) نمایش داده می‌شود. در [۱۹]، از تبدیل فوریه گسسته[۲۲] (DFT) به منظور آموزش واژه‌نامه مختلط استفاده شده تا امکان بازنمایی تُنُک با کیفیت بالا و پیچیدگی محاسباتی کمتری از سیگنال‌های صوتی فراهم شود. در [۲۰]، روش جدیدی برای الگوریتم OMP پیشنهاد شده تا کاربرد آن برای سیگنال‌های گفتاری با ساختار پیچیده بررسی شود. در این روش، واژه‌نامه مختلطی بر روی تبدیل فوریه گسسته مدل گردید و مکانیزم متعامدسازی جدیدی معرفی شد. این اصلاح، توانایی الگوریتم OMP را برای دستیابی به یک بازنمایی فشرده و با کیفیت بالا از سیگنال‌های گفتاری با پیچیدگی محاسباتی کم، افزایش داده است. همچنین یک روش طبقه‌بندی براساس نمونه‌ها و بازنمایی تُنُک در [۲۱] ارائه شده تا از طریق اطلاعات گفتاری، بیماری پارکینسون شناسایی گردد. در این روش، استفاده از واژه‌نامه‌های ویژه هر کلاس به منظور افزایش کارایی روش‌های طبقه‌بندی بازنمایی تُنُک که بر پایه حداقل‌سازی مربعات[۲۳] l_1 و حداقل‌سازی مربعات مثبت[۲۴] (NNLS) می‌باشند، پیشنهاد

[15] Orthogonal Matching Pursuit
[16] K-Singular Value Decomposition
[17] Cluster-Structured Sparse Representation
[18] Voice Conversion
[19] Automatic Speech Recognition
[20] Wavelet Packet Transformation
[21] Sparse Representation Classifier
[22] Discrete Fourier Transform
[23] l1-Regularized Least Squares
[24] Non-Negative Least Squares

شد. هدف، یافتن نمایشی تُنُک برای بردارهای ویژگی‌های گفتاری آزمایشی است که به بازنمایی نمونه‌های گفتاری داده‌های آموزشی ارتباط دارد. یکی از اصلی‌ترین مزایای استفاده از این روش، مقاومت بالای آن در برابر تکرار و نویز داده‌ها و عدم نیاز به تنظیمات پیچیده و زیاد پارامترهای فرابُعدی است.

جدول ۱، گزیده‌ای از پژوهش‌های انجام‌شده در سال‌های اخیر براساس مفاهیم یادگیری واژه‌نامه و بازنمایی تُنُک به منظور پردازش گفتار را نشان می‌دهد. بررسی‌های انجام شده در این جدول بر روی روش‌های ارائه شده براساس ویژگی‌های به کار رفته، اهداف تحقیقاتی، منابع داده‌ای و سال‌های انتشار مقایسه شده است.

در این مقاله به بررسی تعدادی از الگوریتم‌های نمایش تُنُک و رویکردهای یادگیری واژه‌نامه پرداخته می‌شود که نشان‌دهنده ضرورت وجود واژه‌نامه‌های دقیق و متناسب با داده‌ها به منظور پردازش سیگنال گفتار می‌باشد.

در این پژوهش به دنبال توسعه واژه‌نامه‌هایی هستیم که به طور خاص برای بازسازی داده‌های گفتاری بهینه‌سازی شده‌اند. یافته‌ها دراین مقاله عبارتند از:

• ارائه یک بررسی جامع از الگوریتم‌های یادگیری واژه‌نامه برای بازنمایی تُنُک داده‌های گفتاری: این مقاله مروری بر الگوریتم‌های رایج یادگیری واژه‌نامه ارائه می‌دهد.

جدول ۱- پژوهش‌های اخیر در زمینه پردازش گفتار براساس آموزش واژه‌نامه و بازنمایی تُنُک.

مرجع	حوزه ویژگی	هدف	پایگاه داده مورد استفاده	سال انتشار
[۵]	ضرایب کپسترال فرکانس مِل[۲۵] (MFCC)	شناسایی گفتار	TIMIT, AURORA4	۲۰۱۷
[۸]	WPT	بهسازی گفتار	GRID	۲۰۱۷
[۹]	دامنه طیفی-زمانی	VAD	TIMIT, NOISEX_92	۲۰۱۸
[۱۰]	لگاریتم فرکانسی	بهسازی گفتار	NOIZEX_92	۲۰۱۸
[۱۱]	تبدیل فوریه کوتاه مدت[۲۶] (STFT)	بهسازی گفتار	NOIZEUS	۲۰۱۸
[۱۲]	STFT	بهسازی گفتار	NOIZEX_92, GRID	۲۰۱۸
[۱۳]	STFT	بهسازی گفتار	NOIZEX_92, GRID	۲۰۱۸
[۱۴]	حوزه زمان	طبقه‌بندی‌حروف صدادار و بی‌صدا	----	۲۰۱۹
[۱۵]	زمان	بهسازی تبدیل صدا	CMU ARCTIC	۲۰۲۰
[۱۶]	STFT	بهسازی گفتار	NOIZEUS, NOIZEX_92	۲۰۲۰
[۱۷]	تبدیل موجک گسسته[۲۷] (DWT)	بهسازی گفتار	GRID	۲۰۲۰
[۱۸]	WPT	تشخیص خودکار گفتار	TIMIT	۲۰۲۱
[۱۹]	DFT	بازسازی گفتار	TIMIT	۲۰۲۱
[۲۰]	DFT	بازنمایی تُنُک	NOIZEUS, NOIZEX_92	۲۰۲۲
[۲۱]	MFCC	تشخیص بیماری	PC-GITA, MDVR-KCL	۲۰۲۳

• مقایسه عملکرد الگوریتم‌های یادگیری واژه‌نامه در حوزه‌های مختلف: عملکرد الگوریتم‌های یادگیری واژه‌نامه در حوزه‌های زمان، نمایش زمان-فرکانس و تبدیل موجک مورد بررسی قرار می‌گیرد.

• ارائه رویکردی جدید برای آموزش واژه‌نامه‌ها با استفاده از الگوریتم‌های یادگیری واژه‌نامه: این رویکرد می‌تواند به بهبود دقت بازنمایی تُنُک داده‌های گفتاری در طیف وسیعی از اهداف پردازشی کمک نماید.

• ارائه نتایج تجربی که نشان‌دهنده برتری الگوریتم K-SVD در ترکیب با OMP در حوزه STFT است: نتایج

[25] Mel Frequency Cepstral Coefficient
[26] Short-Time Fourier Transform
[27] Discrete Wavelet Trannsform

نشان می‌دهد که این ترکیب، دقت بازسازی سیگنال گفتاری را به طور قابل توجهی افزایش می‌دهد.

• ارائه معیارهای ارزیابی مناسب برای سنجش کارایی الگوریتم‌های یادگیری واژه‌نامه: از معیارهای مختلف زمانی و فرکانسی برای ارزیابی عملکرد الگوریتم‌ها استفاده شده است.

• شناسایی زمینه‌های تحقیقاتی آینده در زمینه یادگیری واژه‌نامه برای بازنمایی تُنُک داده‌های گفتاری: این مقاله زمینه‌های تحقیقاتی بالقوه‌ای را برای بهبود بیشتر این تکنیک‌ها، مانند استفاده از یادگیری عمیق و تکنیک‌های فشرده‌سازی تطبیقی، برجسته می‌نماید.

در بخش دوم مقاله، روش‌های رایج آموزش واژه‌نامه و بازنمایی تُنُک مرور می‌شود. سپس در بخش سوم، الگوریتم پیشنهادی و معیارهای ارزیابی عملکرد آن ارائه می‌گردد. در بخش چهارم به بررسی نتایج بهسازی گفتار در حوزه‌ای که بهترین بازسازی گفتار را داشته، پرداخته می‌شود. در نهایت، در بخش پنجم به بحث در مورد نتایج و جمع‌بندی دستاوردهای تحقیق حاضر پرداخته می‌شود.

۲- یادگیری واژه‌نامه براساس بازنمایی تُنُک

به منظور بهینه‌سازی روال بازنمایی تُنُک و حصول خطای بازنمایی کمتر، ضروری است که واژه‌نامه متناسب با داده‌های مورد بررسی در دسترس باشد. واژه‌نامه‌هایی که مورد استفاده قرار می‌گیرند ممکن است به صورت مجموعه‌ای از توابع پیش آماده مانند تبدیل موجک، تبدیل فوریه، تبدیل گسسته کسینوسی[۲۸] (DCT)، تبدیل Gabor و غیره ساخته شوند. این واژه‌نامه‌ها به دلیل محدودیت‌های خود ممکن است قادر به ارائه نمایش تُنُک برای طیف گسترده‌ای از سیگنال‌ها نباشند. در این صورت واژه‌نامه‌ها می‌توانند براساس داده‌ها و الگوریتم‌های یادگیری آن آموزش داده شوند تا به نحو اثربخشی با داده‌های مورد مطالعه تطبیق پیدا نمایند. گام آغازین در یک الگوریتم یادگیری واژه‌نامه استفاده از یک حوزه تبدیل و سپس انتقال سیگنال گفتاری به آن فضای ویژگی جدید به منظور تحلیل مناسب می‌باشد. این حوزه در روش پیشنهادی، فضای تبدیل موجک، زمان و نمایش زمان-فرکانس در نظر گرفته شده که به صورت زیر قابل نمایش است:

$$S = DX \qquad (۱)$$

در این رابطه S سیگنال گفتار، D واژه‌نامه مورد انتظار و X بیانگر ضرایب در بازنمایی تُنُک می‌باشد. ماتریس داده ورودی $Y \in \mathbb{R}^{P\times F}$ می‌تواند با ترکیب خطی تُنُکی از اتم‌ها 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 $\left\|d_{(:,l)}\right\|_2 = 1, \forall\, l = 1, \dots, L$ و بردار کدگذار K-تنک[۲۹] $X \in \mathbb{R}^{L\times F}, L \gg K$ شامل ضرایب بازنمایی خواهد بود[۲۲]. در حالت کلی مسئله بازنمایی تُنُک که شامل بخش‌های خطای بازسازی و قید تُنُکی می‌باشد به صورت زیر قابل بیان است[۲۳]:

$$X^* = \underset{X}{\operatorname{argmin}} \|Y - DX\|_F^2 , \|X\|_0 \le K \qquad (۲)$$

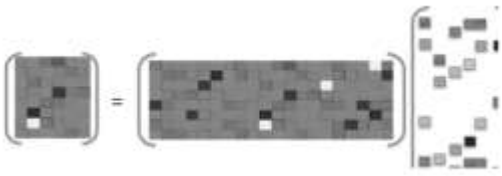
تعداد ضرایب غیرصفر یا کاربنالیتی[۳۰] سیگنال در ماتریس ضرایب تنک X به صورت $\|X\|_0 = K$ خواهد بود. نُرم l_0 در رابطه (۲) منجر به یک مسئله بهینه‌سازی غیرمحدب و NP-Hard می‌شود که ممکن است باعث گیر افتادن در مینیمم‌های محلی شود. جایگزینی نُرم l_0 با نُرم l_1 به عنوان یک روش آرام‌سازی محدب، حل مسئله را تسهیل کرده و احتمال گرفتار شدن در مینیمم‌های محلی را کاهش می‌دهد، اگرچه این جایگزینی لزوماً مسئله را کاملاً محدب نمی‌کند[۱۸].

$$X^* = \underset{X}{\operatorname{argmin}} \|Y - DX\|_F^2 , \|X\|_1 \le \tau \qquad (۳)$$

شکل (۱) نمایشی از نحوه بازنمایی تُنُک سیگنال‌ها را براساس چینش اتم‌ها نمایش می‌دهد.

در مسائل بهینه‌سازی بازنمایی تُنُک، از نرم‌های مختلفی مانند نرم l_0، نرم l_1، نرم l_2 و نرم فروبینیوس استفاده می‌شود. نرم l_0 تعداد عناصر غیرصفر را شمارش می‌کند اما به دلیل غیرمحدب بودن، محاسبات پیچیده‌ای دارد. به همین دلیل، نرم l_1 به عنوان جایگزین محدب برای l_0 استفاده می‌شود تا از گیر افتادن در مینیمم‌های محلی جلوگیری شود. نرم l_2 بیشتر در مسائل مبتنی بر حداقل مربعات و برای کاهش حساسیت به نویز به کار می‌رود. همچنین، نرم فروبینیوس در مسائل مرتبط با ماتریس‌ها، مشابه نرم l_2 برای ماتریس‌ها، برای سنجش اختلاف بین داده‌ها و بازسازی‌ها استفاده می‌شود.

[28] Discrete Cosine Transform
[29] K-Sparse
[30] Cardinality



شکل ۱- بازنمایی تُنُک سیگنال‌ها به صورت ترکیب خطی اتم‌های واژه‌نامه کدگذاری شده.

مسأله یادگیری واژه‌نامه را می‌توان به عنوان یک مسأله غیرمحدب توصیف نمود که به صورت زیر بیان می‌شود:

$$(D^*, X^*) = \underset{D,X}{\operatorname{argmin}} \|Y - DX\|_F^2 \qquad (۴)$$

برای حل مسأله فوق از روش کمینه‌سازی تناوبی[۳۱] استفاده می‌شود [۲۴]. روش کمینه‌سازی تناوبی شامل دو مرحله است. در گام اول ضرایب تُنُک محاسبه می‌شوند و سپس در گام دوم بروزرسانی واژه‌نامه صورت می‌پذیرد. در مرحله اول با فرض ثابت در نظر گرفتن واژه‌نامه، نمایش تُنُک برای دادگان آموزشی محاسبه و ضرایب تُنُک بدست می‌آید. سپس در مرحله بعد با داشتن ضرایب تُنُک، واژه‌نامه بروزرسانی شده و این روال تا رسیدن به خطای کمینه یا حداکثر تعداد تکرار، ادامه می‌یابد. این روال به صورت زیر تعریف می‌شود:

$$\begin{aligned} X^{(K+1)} &= \underset{X}{\operatorname{argmin}} \left\|Y - D^{(K)}X\right\|_F^2 \\ D^{(K+1)} &= \underset{D}{\operatorname{argmin}} \left\|Y - DX^{(K+1)}\right\|_F^2 \end{aligned} \qquad (۵)$$

تفاوت اکثر روش‌های یادگیری واژه‌نامه مربوط به تفاوت در این دو مرحله، به ویژه مرحله بروزرسانی واژه‌نامه می‌باشد. در سال‌های اخیر الگوریتم‌های زیادی برای تقریب ضرایب تُنُک توسعه یافته‌اند. یکی از پرکاربردترین الگوریتم‌ها OMP می باشد. در این الگوریتم، هر مرحله با انتخاب اتمی انجام می‌شود که بیشترین همبستگی را با خطای باقی‌مانده که به صورت تفاوت سیگنال ورودی و ترکیب خطی اتم‌های قبلی محاسبه می‌گردد، داشته باشد. در مرحله اول از آنجایی که هیچ اتمی انتخاب نشده است، خطای باقی‌مانده معادل خود سیگنال ورودی در نظر گرفته می‌شود. علاوه بر این، معیارهای انتخاب بهینه برای بهترین اتم از خطای باقیمانده در تمام ستون‌های واژه‌نامه براساس حاصل ضرب درونی بردار خطای باقیمانده با هر ستون واژه‌نامه و به دنبال آن محاسبه ضریب همبستگی است. اتمی با بالاترین ضریب همبستگی به عنوان بهترین نماینده برای افزودن به اتم‌های قبلی، انتخاب و خطای باقیمانده بین سیگنال ورودی و تقریب فعلی را به حداقل می‌رساند. شرط توقف این الگوریتم تعیین حداکثر تعداد اتم‌ها در نمایش تُنُک سیگنال یا رسیدن به حداقل خطای بازنمایی از پیش تعیین‌شده می‌باشد.

الگوریتم جهت‌های بهینه[۱] (MOD) یکی از روش‌های پایه در آموزش واژه‌نامه می‌باشد. در مرحله اول با استفاده از واژه‌نامه اولیه و یکی از الگوریتم‌های نمایش تُنُک، ضرایب تُنُک حاصل و از الگوریتم OMP در مرحله نخست یادگیری واژه‌نامه استفاده می‌شود[۲۵]. در مرحله بروزرسانی واژه‌نامه، هدف کاهش خطای بازنمایی کلی، $\|Y - DX\|_F^2$ است. برای این منظور، کافی است با استفاده از ماتریس ضرایب تنک به‌دست‌آمده در مرحله‌ی قبل، مشتق خطای بازنمایی نسبت به D محاسبه و برابر صفر قرار داده شود. در نتیجه، یک رابطه برای واژه‌نامه‌ی به‌روزشده‌ی D حاصل می‌شود:

$$D = YX^T(XX^T)^{-1} \qquad (۶)$$

پس از بروزرسانی واژه‌نامه، باید اتم‌های حاصل نرمالیزه شوند. همانطور که مشاهده می‌شود کل مجموعه اتم‌ها در این روش یک بار بروزرسانی می‌شوند که باعث سادگی پیاده‌سازی و سرعت اجرای الگوریتم می‌گردد. روش K-SVD یک از الگوریتم‌های کارآمد و مطرح در زمینه آموزش واژه‌نامه می‌باشد. تفاوت اصلی این روش با روال مبتنی‌بر MOD، در مرحله بروزرسانی واژه‌نامه است. در هر بار تکرار این الگوریتم برخلاف MOD، اتم‌ها به صورت مجزا و همچنین سطر متناظر با این اتم در ماتریس ضرایب تُنُک به طور متوالی بروز می‌شود. در این الگوریتم برای بروزکردن ستون kام واژه‌نامه d_k، سایر اتم‌ها ثابت در نظر گرفته می‌شوند و خطای باقیمانده مرتبط با هر اتم E_k، از رابطه زیر محاسبه می‌شود [۲۶]:

$$\begin{aligned} \|Y - DX\|_F^2 &= \left\|Y - \sum_{i=1}^{L} d_i x^{[i]}\right\|_F^2 \\ &\left\|Y - \sum_{i\neq k}^{L} d_i x^{[i]} - d_k x^{[k]}\right\|_F^2 = \\ &\left\|E_k - d_k x^{[k]}\right\|_F^2 \end{aligned} \qquad (۷)$$

[31] Alternating Minimization

در این رابطه $x^{[i]}$، سطر iام ماتریس X است. با اعمال تجزیه SVD بر روی E_k، اتم d_k ضرایب تُنُک مرتبط با اتم kام $x^{[k]}$ بروز می‌شود. این اتم‌های بروز شده، برای بروزرسانی سایر اتم‌های باقیمانده نیز استفاده می‌شوند. الگوریتم RAMC[۳۲] یکی دیگر از روش‌های آموزش واژه‌نامه است که در سال‌های اخیر بر روی داده‌های آزمایشی تصادفی به خوبی پیاده‌سازی شده است[۲۷]. این روش کاهش همزمان همدوسی متقابل[۳۳] و میانگین همدوسی[۳۴] اتم‌های واژه‌نامه را نتیجه می‌دهد. این دو پارامتر به عنوان حداکثر اندازه همبستگی متقابل بین اتم‌ها $\mu(D)$ و میانگین اندازه همبستگی متقابل بین اتم‌ها $\mu_{avg}(D)$ تعریف شده و طبق رابطه زیر محاسبه می‌شوند[۲۷]:

$$\mu(D) = \max_{i\neq j} |D^T D|_{i,j}$$
$$\mu_{avg}(D) = \sqrt{\frac{\|D^T D - I\|_F^2}{L(L-1)}} \qquad (۸)$$

I ماتریس یکه واحد و L تعداد اتم‌های واژه نامه را مشخص می‌نماید. گام بروزرسانی واژه‌نامه طبق رابطه زیر از طریق گرادیان نزولی محاسبه می‌شود[۲۷]:

$$D^{(k+1)} = D^{(k)} - \alpha \nabla_D F\left(D^{(k)}, X^{(k+1)}\right) \qquad (۹)$$

در این رابطه، F تابع هزینه و $\nabla_D F$،گرادیان آن نسبت به D را نشان می‌دهد. همچنین α، شیب گرادیان نزولی را تنظیم می‌نماید[۲۷]. این روش نیز در هر مرحله بروزرسانی واژه‌نامه همانند MOD تمامی اتم‌ها را بروز می‌نماید. به دلیل ضرب داخلی D و X، آموزش واژه‌نامه یک مسئله غیرمحدب است. در [۲۸]، این روش به صورت یک مسئله محدب ارائه شده که با جایگزینی عبارت محدب $X = X_0 + (X - X_0)$ و $D = D_0 + (D - D_0)$ در رابطه ۴ آموزش واژه‌نامه جدیدی طبق رابطه ۱۰ بیان می‌شود:

$$(D^*, X^*) = \underset{D,X}{\operatorname{argmin}} \|Y + D_0 X_0 - (D_0 X + D X_0)\|_F^2 \qquad (۱۰)$$

با ثابت نگاه‌داشتن پارامترهای مختلف در هنگام کمینه‌سازی تناوبی D و X، چهار الگوریتم حاصل می‌شود. یکی از این الگوریتم‌ها که سرعت و دقت آن بر روی داده‌های مصنوعی نسبت به سایر الگوریتم‌های معرفی شده برتری دارد، کمینه‌سازی تناوبی به صورت رابطه زیر است:

$$R^{(k)} = Y - \left(D^{(k)} - D^{(k-1)}\right)X^{(k)}$$
$$X^{(k+1)} = \underset{X}{\operatorname{argmin}} \left\|R^{(k)} - D^{(k-1)}X\right\|_F^2 \qquad (۱۱)$$
$$D^{(k+1)} = \underset{D}{\operatorname{argmin}} \left\|Y - DX^{(k+1)}\right\|_F^2$$

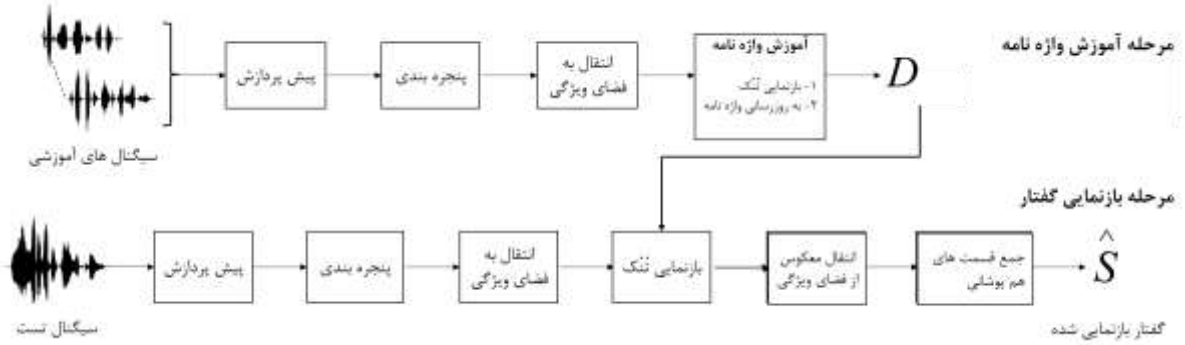
گام بروزرسانی واژه‌نامه در این روش همان روال مورد استفاده در MOD می‌باشد. این روش با نام اختصاری UD4-MOD[۳۵] در [۲۸] معرفی شده است.

۳- بازنمایی سیگنال گفتار مبتنی‌بر یادگیری واژه‌نامه

سیگنال گفتار، اطلاعات پیچیده‌ای را در مورد گوینده، لحن و محتوای گفتار به همراه دارد. یکی از راه‌های پردازش و تحلیل این اطلاعات، روش‌های مبتنی‌بر بازنمایی سیگنال گفتار به صورت فشرده و دقیق است. به این منظور دسترسی به واژه‌نامه‌ای که اطلاعات دقیق و برجسته سیگنال گفتار در فضای ویژگی را مدنظر قرار دهد، ضروری است. در این بخش، به بررسی روش‌های مبتنی‌بر یادگیری واژه‌نامه برای بازنمایی سیگنال گفتار پرداخته می‌شود. این روش‌ها، واژه‌نامه را از خود داده‌های گفتاری می‌آموزد و از آن برای بازنمایی سیگنال گفتار به صورت ترکیبی خطی از اتم‌ها استفاده می‌کند.

در شکل (۲)، بلوک دیاگرام مرحله آموزش و بازنمایی سیگنال گفتار در فضای ویژگی مورد استفاده، نشان داده شده است. برای ارزیابی کارایی بازنمایی از معیارهای حداقل میانگین خطا[۳۶] (MSE)، خطای بازسازی[۳۷](RE)، نسبت سیگنال به نویز بخش‌بندی شده[۳۸] (SegSNR)، (fwSegSNR)[۳۹]، ارزیابی ادراکی کیفیت گفتار[۴۰] (PESQ) و درک عینی کوتاه مدت[۴۱] (STOI) استفاده شده است. فضاهای ویژگی متداول مورد استفاده برای آموزش واژه‌نامه با توجه به کاربردهای مختلف سیگنال گفتار، حوزه‌های زمان، نمایش زمان-فرکانس و تبدیل بسته موجک و ... می‌باشد. واژه‌نامه گفتار در حوزه‌های ذکرشده براساس بازنمایی سیگنال‌های گفتار آموزش داده می‌شوند.

[32] Reduced Average and Mutual Coherence
[33] Mutual Coherence
[34] Average Coherence
[35] UpDated4-MOD
[36] Mean Square Error
[37] Reconstruction Error
[38] Segmental SNR
[39] Ferquency WSegmental SNR
[40] Perceptual Evaluation of Speech Quality
[41] Short Time Objective Intelligibility



شکل ۲- بلوک دیاگرام مرحله آموزش واژه‌نامه و بازنمایی گفتار.

به منظور بررسی روال بازنمایی، از گفتار پایگاه داده NOISEUES با فرکانس نمونه‌برداری ۸ کیلوهرتز در تمامی حوزه‌ها استفاده شده است. در ابتدا فضای ویژگی مطابق با روال بررسی شده در [۱۵] و [۱۴] در حوزه زمان در نظر گرفته می‌شود. با توجه به بلوک دیاگرام در شکل (۲)، گفتار در مرحله آموزش با استفاده از پنجره‌های همینگ با طول ۱۲/۵ میلی‌ثانیه و همپوشانی ٪۵۰ تقسیم‌بندی می‌شود تا داده‌های آموزشی برای آموزش واژه‌نامه در این حوزه ایجاد گردد. نرخ افزونگی واژه‌نامه‌های فراکامل به صورت تجربی و با توجه به کمترین خطای بازسازی در شبیه‌سازی‌ها ، مقدار ۴ در نظر گرفته شده است [۸]. انتخاب این پارامتر به مقداری بیشتر از این مقدار، ضمن بالابردن زمان و هزینه محاسبات، تاثیر قابل توجهی در بهبود نتایج شبیه‌سازی ندارد. همچنین، انتخاب عددی کمتر از این مقدار برای نرخ افزونگی، کاهش کیفیت و افزایش خطای بازنمایی را نتیجه می‌دهد. این پارامتر با در نظر گرفتن توازنی بین زمان شبیه‌سازی و دقت نتایج بدست آمده، انتخاب شده است. شبیه‌سازی‌های انجام شده برای تعیین نرخ افزونگی براساس فضاهای ویژگی مختلف بکارگرفته شده و نیز روش آموزش واژه‌نامه K-SVD در جدول ۲ گزارش شده است. در جدول ۳، نتایج حاصل از آزمایش‌های مختلف در حوزه زمان با روش‌های مختلف آموزش MOD، K-SVD، RAM و UD4-MOD گزارش شده است. در [۸] و [۱۸]، WPT به عنوان فضای ویژگی برای آموزش واژه‌نامه بکار گرفته شده است. در این بخش، دو سناریو برای بازنمایی گفتار با WPT مورد بررسی قرار گرفته‌اند. در روش اول مطابق شکل (۲)، سیگنال گفتار پس از پنجره‌گذاری، تا سطح ۳ با WPT تجزیه شده و هر قاب به ۸ زیرباند ضرایب بسته موجک تجزیه می‌شود. با داشتن این ۸ زیرباند ضرایب، ۸ واژه‌نامه مرتبط برای هر زیرباند آموزش داده می‌شود. در سناریو دوم، قبل از اعمال بلوک پنجره‌گذاری، WPT از گفتار کوتاه‌مدت حاصل و سپس، قاب‌بندی روی زیرباندهای ضرایب بسته موجک اعمال می‌گردد تا داده‌ها برای آموزش واژه‌نامه‌های مرتبط با هر زیرباند در دسترس باشند. در ادامه، ۸ واژه‌نامه برای استفاده در مرحله بازنمایی گفتار آموزش داده می‌شوند. جدول ۴ نتایج شبیه‌سازی‌ها برای دو سناریو فوق در فضای WPT را نشان می‌دهد. استفاده از حوزه STFT به دلیل توانایی تحلیل و نمایش همزمان حوزه زمان و فرکانس سیگنال‌ها، یکی از انتخاب‌های پرکاربرد به منظور آموزش واژه‌نامه گفتار محسوب می‌شود. در پژوهش‌های [۱۱-۱۲]، [۱۳، ۱۶]، آموزش واژه‌نامه گفتار براساس تبدیل فوریه کوتاه‌مدت سیگنال گفتار انجام شده است. برای دستیابی به این هدف، در مرحله‌ی آموزش، از دامنه‌ی حوزه STFT بهره‌برده شده است. این در حالی است که برای بازنمایی گفتار، فاز سیگنال بدون تغییر مورد استفاده قرار گرفته است. در جدول ۵، نتایج شبیه‌سازی‌های STFT گزارش شده که کارایی این روش در آموزش و بازنمایی گفتار را نشان می‌دهد.

جدول ۲- تأثیر نرخ افزونگی بر خطای بازنمایی تُنُک سیگنال گفتار براساس فضاهای ویژگی مختلف و هزینه محاسباتی.

فضای ویژگی	معیارهای ارزیابی	نرخ افزونگی			
		۲	۴	۶	۸
زمان	زمان آموزش	۲۱	۳۲	۴۲	۵۶
	زمان بازنمایی	۱/۰۶	۴/۶	۹/۳	۱۳/۹
	RE	۰/۷۳	۰/۶۱	۰/۵۷	۰/۵۵
	MSE	۲/۵e-۵	۱/۸e-۵	۱/۶e-۵	۱/۵e-۵
WPT1	زمان آموزش	۲۶	۴۲	۴۹	۶۳
	زمان بازنمایی	۱/۳	۲/۲	۲/۶	۷/۶
	RE	۰/۵۷	۰/۵۳	۰/۵۲	۰/۵۱
	MSE	۱/۵۸e-۵	۱/۳۷e-۵	۱/۳۳e-۵	۱/۳۰e-۵
WPT2	زمان آموزش	۲۹	۴۳	۵۸	۷۰
	زمان بازنمایی	۱/۳	۲/۲	۲/۴	۷/۸
	RE	۰/۴۷	۰/۴۴	۰/۴۴	۰/۴۲
	MSE	۱/۱e-۵	۰/۹۴e-۵	۰/۹۲e-۵	۰/۸۶e-۵
STFT	زمان آموزش	۲۱	۳۲	۴۲	۵۴
	زمان بازنمایی	۱/۱	۴/۶	۹/۳	۱۵/۹
	RE	۰/۳۹	۰/۲۶	۰/۲۳	۰/۲۰
	MSE	۰/۷۷e-۵	۰/۳۵e-۵	۰/۲۵e-۵	۰/۱۹e-۵

جدول ۳- نتایج معیارهای ارزیابی بازنمایی گفتار در حوزه زمان.

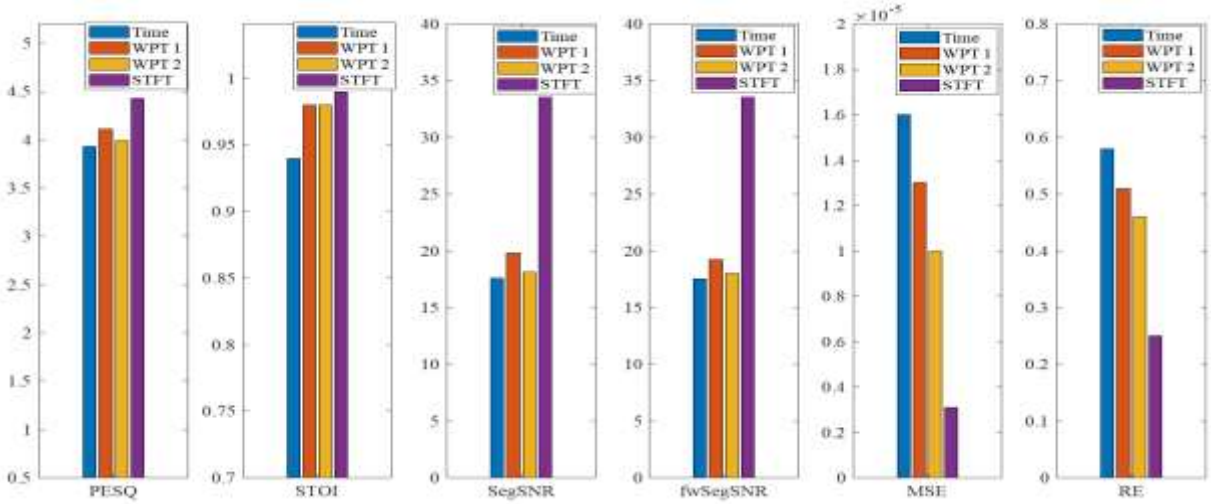
روش آموزش واژه‌نامه	MSE	RE	SegSNR	fwSegSNR	STOI	PESQ
MOD	۱/۸e-۵	۰/۶۲	۱۶/۸۳	۱۶/۸۴	۰/۹۳	۳/۸۷
K-SVD	۱/۷e-۵	۰/۶۰	۱۶/۸۸	۱۶/۸۹	۰/۹۴	۳/۸۷
RAMC	۳/۸e-۵	۰/۸۸	۱۵/۵۹	۱۵/۳۹	۰/۹۲	۳/۲۸
UD4-MOD	۱/۶e-۵	۰/۵۸	۱۷/۶۱	۱۷/۵۲	۰/۹۴	۳/۹۳

جدول ۴- نتایج معیارهای ارزیابی بازنمایی گفتار در حوزه WPT.

روش آموزش واژه‌نامه		MSE	RE	SegSNR	fwSegSNR	STOI	PESQ
MOD	روش ۱	۱/۳e-۵	۰/۵۱	۱۹/۸۳	۱۹/۲۳	۰/۹۸	۴/۱۱
	روش۲	۱ e-۵	۰/۴۶	۱۸/۱۳	۱۸/۰۱	۰/۹۸	۳/۹۹
K-SVD	روش ۱	۱/۴e-۵	۰/۵۴	۱۹/۶۲	۱۸/۹۲	۰/۹۸	۴/۱۰
	روش۲	۱/۴e-۵	۰/۴۶	۱۷/۷۴	۱۷/۶۲	۰/۹۷	۳/۹۴
RAMC	روش ۱	۱/۷e-۵	۰/۵۹	۱۹/۰۶	۱۸/۲۵	۰/۹۸	۳/۹۸
	روش۲	۱/۸e-۵	۰/۶۲	۱۶/۶۱	۱۶/۵۳	۰/۹۶	۳/۸۰
UD4-MOD	روش ۱	۱/۵e-۵	۰/۵۶	۱۸/۷۵	۱۷/۹۹	۰/۹۷	۳/۹۸
	روش۲	۱/۶e-۵	۰/۵۹	۱۶/۰۴	۱۶/۰۴	۰/۹۶	۳/۷۶

جدول ۵- نتایج معیارهای ارزیابی بازنمایی گفتار در حوزه STFT.

روش آموزش واژه‌نامه	MSE	RE	SegSNR	fwSegSNR	STOI	PESQ
MOD	۳/۲e-۶	۰/۲۵	۳۲/۹۸	۳۲/۹۲	۰/۹۸	۴/۱۴
K-SVD	۳/۱e-۶	۰/۲۵	۳۳/۵۶	۳۳/۴۷	۰/۹۹	۴/۴۳
RAMC	۳/۲e-۶	۰/۲۵	۳۳/۴۶	۳۳/۳۹	۰/۹۹	۴/۴۲
UD4-MOD	۷/۹e-۶	۰/۴۰	۳۰/۳۳	۳۰/۲۱	۰/۹۸	۴/۳۴



شکل ۳- مقایسه معیارهای ارزیابی بازنمایی گفتار براساس معیارهای مختلف ارزیابی گفتار در حوزه‌های مختلف بازنمایی.

بررسی جداول ۳، ۴ و ۵ نشان می‌دهد که در حوزه زمان آموزش واژه‌نامه براساس روش UD4-MOD در مقایسه با سایر الگوریتم‌ها کارایی بیشتری دارد. در فضای ویژگی WPT به منظور بازنمایی گفتار، روش ۱ در مقایسه با روش ۲ عملکرد مطلوب‌تری دارد. بر طبق معیارهای ارزیابی، آموزش واژه‌نامه براساس الگوریتم MOD، در مقایسه با سایر روش‌های آموزشی مورد استفاده در این فضا موثرتر عمل می‌کند و همچنین الگوریتم K-SVD مورد استفاده برای آموزش واژه‌نامه در حوزه STFT نسبت به سایر روش‌ها کارایی بالاتری را به نمایش می‌گذارد. به منظور بررسی عمیق‌تر این مسئله که بازنمایی گفتار در کدام حوزه موفق‌تر بوده، نتایج برتر ذکر شده در هر فضای ویژگی برای همه معیارهای ارزیابی به عنوان شاخص آن حوزه در شکل (۳) نمایش داده شده است. همانطور که در نمودارهای این شکل مشاهده می‌شود، تمام معیارها گویای برتری قاطع حوزه STFT نسبت به سایر فضاهای مورد بررسی است. حوزه مربوط به روش اول فضای WPT و روش دوم WPT و در نهایت حوزه زمان در رتبه‌های بعدی قرار می‌گیرند.

۴- بهسازی گفتار براساس بازنمایی گفتار

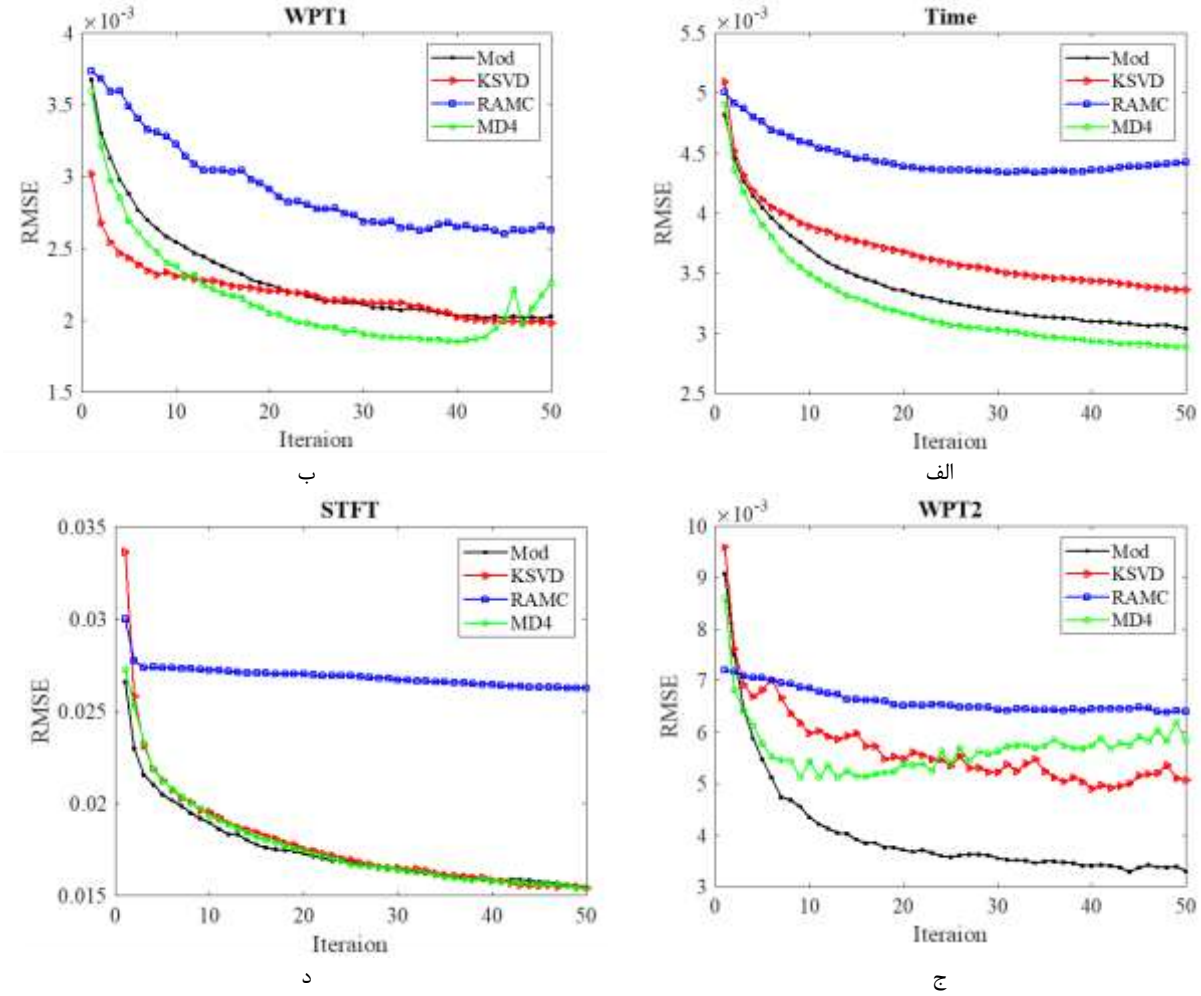
همانطور که بیان شد بهسازی گفتار به عنوان یکی از کاربردهای بازنمایی و یادگیری واژه‌نامه در حوزه پردازش سیگنال شناخته می‌شود. در این بخش با توجه به کارایی بالا، تنها حوزه STFT نسبت به سایر حوزه‌ها در مسئله بازنمایی به عنوان فضای ویژگی در کاربرد بهسازی گفتار در نظر گرفته می‌شود. دامنه STFT سه نوع نویز سفید، ماشین و خیابان همانند گفتار تمیز توسط الگوریتم K-SVD آموزش داده شده است و با کنار هم قرار دادن واژه‌نامه گفتار بدون نویز D_s، واژه‌نامه نویز D_n، یک واژه‌نامه مرکب برای بررسی بهسازی متناسب با هر کدام از نویزها به صورت مجزا به صورت $D = [D_s, D_n]$ تشکیل می‌شود. سپس سیگنال‌های نویز با SNRهای ۰، ۵ و ۱۰ دسی‌بل به گفتار بدون نویز اضافه شده و مسئله بازنمایی با الگوریتم OMP با گفتار نویزی و واژه‌نامه مرکب بر طبق رابطه ۱۲ و ۱۳ دنبال می‌شود [۱۲]:

$$S_n \simeq DX = [D_s, D_n]\begin{bmatrix} X_s \\ X_n \end{bmatrix}$$

$$\hat{X} = \underset{X}{\operatorname{argmin}} \left\| S_n - [D_s, D_n]\begin{bmatrix} X_s \\ X_n \end{bmatrix} \right\|_F^2 \qquad (۱۲)$$

$$\hat{X} = OMP(S_n\ , [D_s, D_n])$$

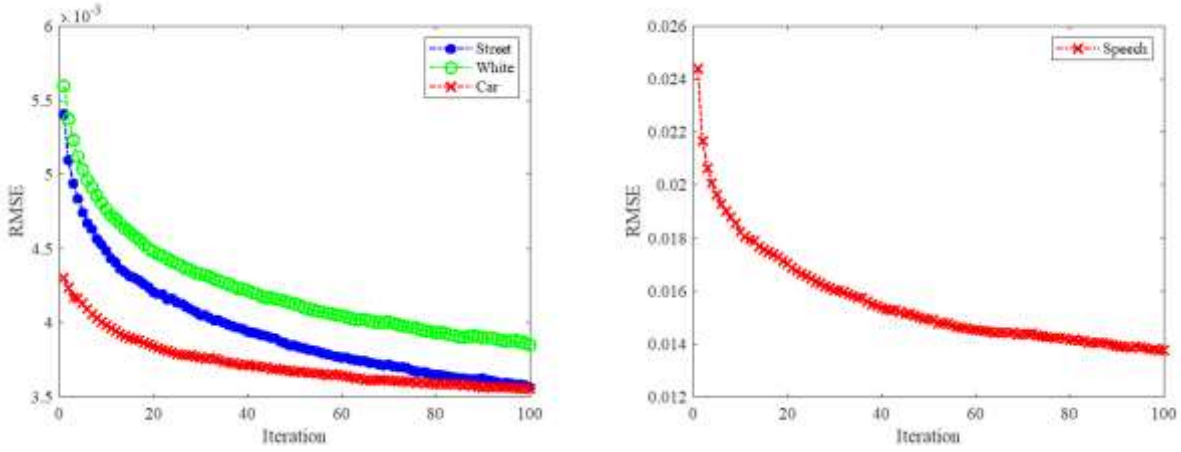
$$\hat{S}_s = D_s\hat{X}_s\ , \hat{N} = D_n\hat{X}_n \qquad (۱۳)$$



شکل ۴- منحنی‌های RMSE مرحله آموزش گفتار در حوزه، الف) زمان، ب) WPT-روش اول، ج) WPT-روش دوم، د)STFT.

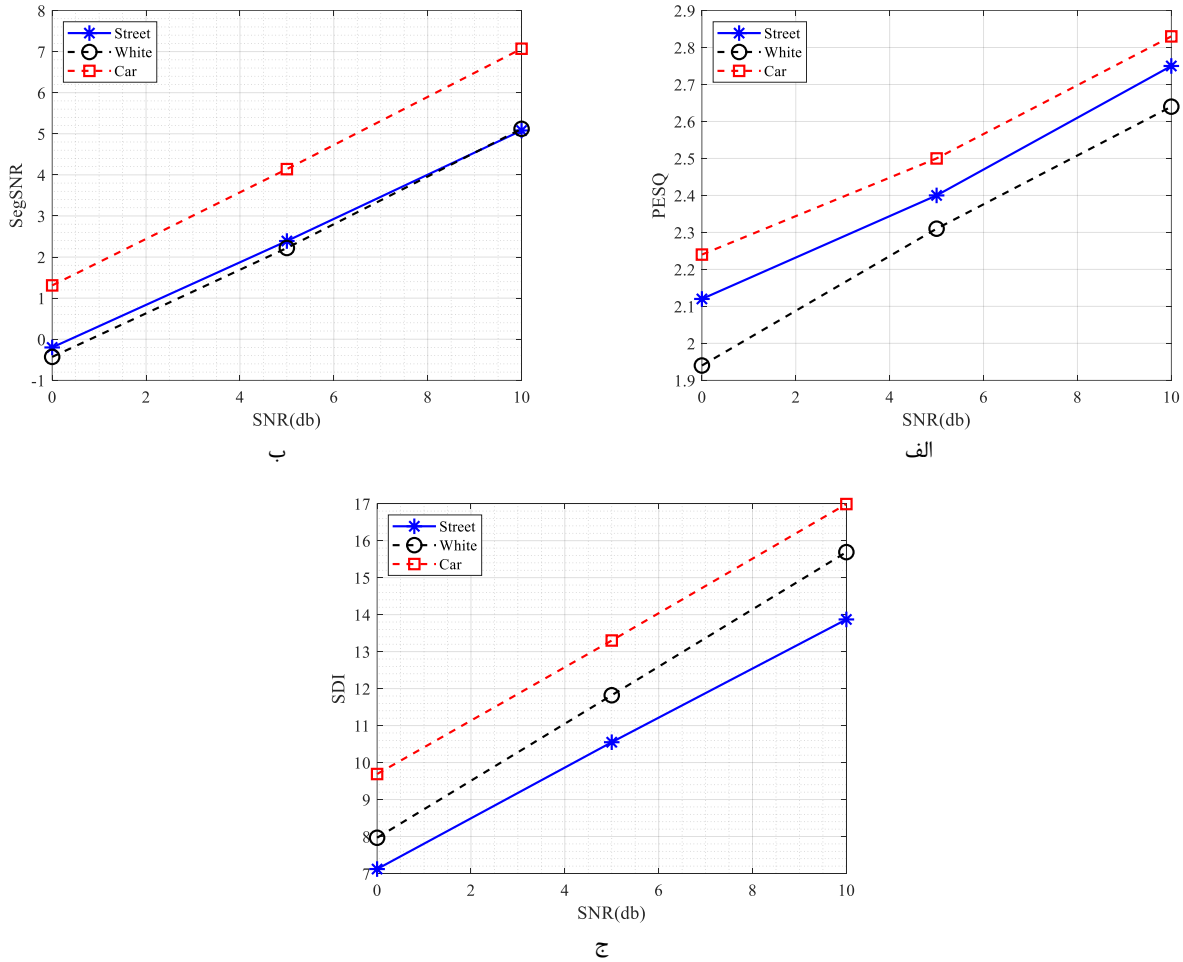
در روابط ۱۲ و ۱۳، S_n، دامنه طیف گفتار نویزی، به صورت ترکیبی تُنُک از واژه‌نامه گفتار تمیز و نویز در نظر گرفته شده است.

در مرحله بهسازی تلاش بر جداسازی اجزای گفتار تمیز و نویز با استفاده از واژه‌نامه مرکب است. X_s، X_n زیرماتریس‌های از X هستند که باید تخمین زده شوند. این کار با کمک الگوریتم بازنمایی OMP انجام می‌شود. $\hat{X}_s$ و $\hat{X}_n$ به ترتیب تخمینی از ماتریس ضرایب تُنُک برای بازنمایی حدی اندازه طیف گفتار تمیز $\hat{S}_s$، و تخمین اندازه طیف نویز $\hat{N}$، نسبت به واژه‌نامه‌های مرتبط با آن‌ها یعنی D_s و D_n هستند. فیلتری شبیه فیلتر وینر بر روی دامنه‌های طیفی بدست آمده اعمال می‌شود و در نهایت با ترکیب فاز گفتار نویزی و خروجی حاصل از فیلتر، گفتار بهسازی شده حاصل می‌شود. منحنی‌های RMSE مرحله آموزش گفتار در حوزه‌های مختلف، زمان، WPT-روش اول، WPT-روش دوم، و STFT در شکل (۴) نمایش داده شده است. همچنین شکل (۵)، منحنی‌های RMSE بازسازی واژه‌نامه برای گفتار تمیز و نویز با الگوریتم آموزشی K-SVD را نشان می‌دهد. نرخ افزونگی در این آزمایش برای آموزش واژه‌نامه‌ها بر طبق نتایج گزارش شده در جدول ۲، برابر ۴ و مرحله تکرار الگوریتم ۱۰۰ در نظر گرفته شده است. در این مقاله سه معیار PESQ، SegSNR و SDI، برای ارزیابی کیفیت گفتار بهسازی شده با شرایط نویزی متفاوت در نظر گرفته شده‌است. نتایج بهسازی شده در نمودارهای شکل (۶) نشان داده شده‌است.



ب الف

شکل ۵- منحنی‌های RMSE مرحله آموزش الف)گفتار تمیز، ب) نویز، با الگوریتم KSVD در حوزه STFT.



ب الف

ج

شکل۶- نتایج بهسازی گفتار با معیارهای الف) PESQ، ب) SegSNR و ج)شاخص اعوجاج گفتار[۴۲] (SDI).

[42] Speech Distrotion Index

۶- نتیجه‌گیری

در این مقاله، روش‌های مختلف یادگیری واژه‌نامه و بازنمایی تُنُک به منظور بازسازی سیگنال‌های گفتاری مورد بررسی قرار گرفته است. در ابتدا، به طور خلاصه به مروری بر کاربردهای آموزش واژه‌نامه و بازنمایی تُنُک در پردازش سیگنال، به طور خاص سیگنال گفتار پرداخته شده است. سپس، الگوریتم‌های رایج یادگیری واژه‌نامه از جمله MOD و K-SVD، همچنین الگوریتم‌های جدید و کارآمد مانند RAMC و UD4-MOD که کارایی آن‌ها تاکنون با مجموعه دادگان مصنوعی آزمایش شده بودند معرفی و مورد بحث قرار گرفت. نتایج حاصل از شبیه‌سازی‌ها با معیارهای RE، MSE، SegSNR، fwSegSNR، PESQ و STOI نشان داد که استفاده از الگوریتم یادگیری واژه‌نامه K-SVD در ترکیب با الگوریتم بازنمایی تُنُک در حوزه STFT، بهترین نتایج بازسازی گفتار در مقایسه با سایر الگوریتم‌ها را نشان می‌دهد. این الگوریتم نتایج مطلوبی برای بازسازی سیگنال گفتار در سایر حوزه‌ها نیز نشان داده است. انتخاب الگوریتم مناسب برای یادگیری واژه‌نامه در بازنمایی تُنُک سیگنال‌های گفتاری، به عواملی مانند نوع داده، دقت مورد نظر، زمان آموزش و کاربرد نهایی بازنمایی بستگی دارد. با این حال، الگوریتم K-SVD به طور کلی به عنوان روشی کارآمد و دقیق برای بازنمایی تُنُک سیگنال‌های گفتاری شناخته شده است. در نهایت، به عنوان یکی از کاربردهای بازنمایی گفتار، به بررسی بهبود کیفیت گفتار در شرایط نویزی در حوزه STFT پرداخته شد. نتایج شبیه‌سازی با معیارهای PESQ، SegSNR و SDI نشان‌دهنده کارایی بالای این روش با آموزش واژه‌نامه و بازنمایی تُنُک می‌باشد. به عنوان پیشنهاداتی برای تحقیقات آینده، طراحی الگوریتم‌های یادگیری واژه‌نامه ترکیبی، بررسی کاربرد بازنمایی گفتار در سایر حوزه‌ها، و توسعه روش‌های یادگیری واژه‌نامه حوزه‌های مختلف پردازش سیگنال پیشنهاد می‌شود.

تقدیر و تشکر

نویسندگان از تمامی افرادی که به‌طور غیرمستقیم در انجام این پژوهش همکاری و راهنمایی داشته‌اند، قدردانی می‌کنند.

تعارض منافع

نویسندگان اعلام می‌کنند که هیچ‌گونه تعارض منافع در مورد انتشار این مقاله وجود ندارد.

تأییدیه اخلاقی

نویسندگان متعهد می‌شوند که مطالب این مقاله پیش از این در هیچ مجله‌ای منتشر نشده و همزمان برای بررسی و چاپ به نشریه دیگری ارسال نشده است.

مشارکت نویسندگان

ناصر شرفی: طراحی و پیاده‌سازی الگوریتم‌ها، انجام شبیه‌سازی‌ها، تحلیل نتایج و نگارش اولیه مقاله.

سلمان کریمی: ارائه ایده پژوهش، نظارت علمی بر روند تحقیق، بازبینی روش‌شناسی و اصلاح و ویرایش علمی مقاله.

سمیرا مودتی: مشارکت در ایده پژوهش و تحلیل نتایج، ارائه مشاوره تخصصی، بازبینی فنی و علمی مقاله و تأیید نسخه نهایی.

منابع مالی

این پژوهش هیچ‌گونه حمایت مالی از سازمان‌ها، نهادها یا مؤسسات دولتی و خصوصی دریافت نکرده است.

مراجع

[1] Tsaig, Yaakov, and David L. Donoho. "Extensions of Compressed Sensing." Signal Processing 86, no. 3 (2006): 549–571.

[2] Zhao, Yongqiang, and Jingxiang Yang. "Hyperspectral Image Denoising via Sparse Representation and Low-Rank Constraint." IEEE Transactions on Geoscience and Remote Sensing 53, no. 1 (2015): 296–308.

[3] Yang, Meng, Dengxin Dai, Linlin Shen, and Luc Van Gool. "Latent Dictionary Learning for Sparse Representation Based Classification." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 4138–4145. 2014.

[4] Liu, Yu, Xun Chen, Aiping Liu, Rabab K. Ward, and Z. Jane Wang. "Recent Advances in Sparse Representation Based Medical Image Fusion." IEEE Instrumentation & Measurement Magazine 24, no. 2 (2021): 45–53.

[5] Sharma, Pulkit, Vinayak Abrol, and Anil Kumar Sao. "Deep-Sparse-Representation-Based Features for Speech Recognition." IEEE/ACM Transactions on Audio, Speech, and Language Processing 25, no. 11 (2017): 2162–2175.

[6] Peng, Peng, Yongfeng Ju, Yipu Zhang, Kaiming Wang, Suying Jiang, and Yuping Wang. "Sparse Representation and Dictionary Learning Model Incorporating Group Sparsity and Incoherence to Extract Abnormal Brain Regions Associated with Schizophrenia." IEEE Access 8 (2020): 104396–104406.

[7] Tourneret, Jean-Yves, Adrian Basarab, Nora Leïla Ouzir, and Qi Wei. "Sparse Representations and Dictionary Learning: From Image Fusion to Motion Estimation." In 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, 25–28. 2021.

[8] Mavaddaty, Samira, Seyed Mohammad Ahadi, and Sanaz Seyedin. "Speech Enhancement Using Sparse Dictionary Learning in Wavelet Packet Transform Domain." Computer Speech & Language 44 (2017): 22–47.

[9] Eshaghi, Mohadese, Farbod Razzazi, and Alireza Behrad. "A Voice Activity Detection Algorithm in Spectro-Temporal Domain Using Sparse Representation." International Journal of Machine Learning and Cybernetics 10, no. 7 (2019): 1791–1803.

[10] Sugiura, Yosuke, and Tetsuya Shimamura. "Speech Enhancement Based on Sparse Representation in Logarithmic Frequency Scale." In 2018 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), 252–257. 2018.

[11] Shaheen, Dima, Oumayma Al Dakkak, and Mohiedin Wainakh. "Incoherent Discriminative Dictionary Learning for Speech Enhancement." Journal of Telecommunications and Information Technology (2018): 42–54.

[12] Ji, Yunyun, Wei-Ping Zhu, and Benoit Champagne. "Speech Enhancement Based on Dictionary Learning and Low-Rank Matrix Decomposition." IEEE Access 7 (2019): 4936–4947.

[13] Bai, Huang, Chuanrong Hong, Sheng Li, Yimin D. Zhang, and Xiumei Li. "Unit-Norm Tight Frame-Based Sparse Representation with Application to Speech Inpainting." Digital Signal Processing 123 (2022): 103426.

[14] Wang, Lianzi, Nikos Mastorakis, and Xiaodong Zhuang. "Voiced/Unvoiced Pronunciation Judgement Based on Sparse Representation and Learning Dictionary." In MATEC Web of Conferences 292 (2019): 04012.

[15] Ding, Shaojin, Guanlong Zhao, Christopher Liberatore, and Ricardo Gutierrez-Osuna. "Learning Structured Sparse Representations for Voice Conversion." IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2020): 343–354.

[16] Haneche, Houria, Bachir Boudraa, and Abdeldjalil Ouahabi. "A New Way to Enhance Speech Signal Based on Compressed Sensing." Measurement 151 (2020): 107236.

[17] Mavaddati, Samira. "A New Method for Speech Enhancement Based on Incoherent Model Learning in Wavelet Transform Domain." Signal and Data Processing 17, no. 3 (2020): 17–36.

[18] Al-Hassani, Ihsan, Oumayma Al-Dakkak, and Abdlnaser Assami. "Phonetic Segmentation Using a Wavelet-Based Speech Cepstral Features and Sparse Representation Classifier." Journal of Telecommunications and Information Technology 4 (2021): 12–22.

[19] Kwek, Lee-Chung, Alan W. C. Tan, Heng-Siong Lim, Cheah Heng Tan, and Khaled Abdulaziz Alaghbari. "Sparse Representation and Reproduction of Speech Signals in Complex Fourier Basis." International Journal of Speech Technology 25, no. 1 (2021): 211–217.

[20] Sun, Linhui, Yunyi Bu, Pingan Li, and Zihao Wu. "Single-Channel Speech Enhancement Based on Joint Constrained Dictionary Learning." EURASIP Journal on Audio, Speech, and Music Processing 2021, no. 29 (2021): 1–14.

[21] Reddy, M. Kiran, and Paavo Alku. "Exemplar-Based Sparse Representations for Detection of Parkinson's Disease from Speech." IEEE/ACM Transactions on Audio, Speech, and Language Processing 31 (2023): 1386–1396.

[22] Xie, Yuan, Kan Xie, and Shengli Xie. "Underdetermined Blind Source Separation of Speech Mixtures Unifying Dictionary Learning and Sparse Representation." International Journal of Machine Learning and Cybernetics 12, no. 12 (2021): 3573–3583.

[23] Chen, Xiaoju, Y. Li, S. Ding, B. Tan, and Y. Jiang. "A Novel Nonlinear Dictionary Learning Algorithm Based on Nonlinear-KSVD and Nonlinear-MOD." In Lecture Notes in Computer Science, 167–179. 2022.

[24] Nejati, Mansour. "Improving the Performance of Sparse Representation Model for Image Restoration and Reconstruction." PhD diss., Isfahan University of Technology, 2016.

[25] Cai, Shuting, Shaojia Weng, Binling Luo, Daolin Hu, Simin Yu, and Shuqiong Xu. "A Dictionary-Learning Algorithm Based on Method of Optimal Directions and Approximate K-SVD." In 2016 35th Chinese Control Conference (CCC), 4793–4798. 2016.

[26] Deeba, Farah, and Kun She. "Lossless Digital Image Watermarking in Sparse Domain by Using K-Singular Value Decomposition Algorithm." IET Image Processing 14, no. 10 (2020): 2080–2087.

[27] Parsa, Javad, Mostafa Sadeghi, Massoud Babaie-Zadeh, and Christian Jutten. "Joint Low Mutual and Average Coherence Dictionary Learning." In Proceedings of the 26th European Signal Processing Conference (EUSIPCO 2018), 1739–1743. 2018.

[28] Parsa, Javad, Mostafa Sadeghi, Massoud Babaie-Zadeh, and Christian Jutten. "A New Algorithm for Dictionary Learning Based on Convex Approximation." In Proceedings of the 27th European Signal Processing Conference (EUSIPCO 2019). 2019.