2023 10th International Conference on Signal Processing and Integrated Networks (SPIN) | 978-1-6654-9099-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/SPIN57001.2023.10116123

# A Comparative Analysis of Speech Enhancement Techniques Based on Sparsity Features

Raj Kumar
Electrical Engineering Department
Indian Institute of Technology
Roorkee, India
rajkciitr@gmail.com

Manoj Tripathy
Electrical Engineering Department
Indian Institute of Technology
Roorkee, India
monoj.tripathy@ee.iitr.ac.in

R. S. Anand
Electrical Engineering Department
Indian Institute of Technology
Roorkee, India
r.anand@ee.iitr.ac.in

***Abstract*— This paper presents a comparative analysis of speech enhancement algorithms based on speech sparsity features. The sparsity of a speech is a signal-specific characteristic that plays a crucial role in identifying speech signal components from noisy speech. Conventional enhancement techniques like Spectral subtraction, Sub-space and statistical methods require knowledge of noise distribution for better performance. The paper compares major sparsity-based enhancement by highlighting the merits and demerits by comparing quality and intelligibility. An objective evaluation of quality and intelligibility has been performed to show relative performance on a common database. Results show that dictionary-based speech enhancement aided with voice activity detection and spectral subtraction performs better than others. The performances improve further if noise-type and noise dictionary is known as apriori in the joint dictionary and non-negative matrix factorization method. This analysis will help the researcher understand sparsity-based enhancement techniques and develop techniques to overcome existing issues.**

***Keywords*— *Speech Enhancement, Sparsity, Intelligibility, Compressive Sensing Theory***

## I. INTRODUCTION

Smartphones, voice-based devices, etc., require good-quality input speech for better performance, which is affected by environmental noise. Speech enhancement aims to recover the speech signal from the mixture of clean speech and additive noise signals. The performance of speech enhancement is measured in terms of noise suppression related to quality and extent of distortion related to speech intelligibility. It is a challenging task to maximize quality with less distortion. There exist a trade-off between SNR gain and intelligibility improvement [1]. Naturally, speech and noise are highly nonstationary signals. Conventional Speech Enhancement (SE) techniques [2] rely on noise estimation, e.g., in the case of spectral subtraction method [3] or noise power tracking [4], [5] or parameter of noise model in case of statistical method [6]. However, they exploit very less information from speech; hence their performance is poor in case of non-stationary noise and under very low SNR conditions [7]. Noise estimation is a very tough task under heavy and non-stationary noisy conditions. In recent times, speech sparsity has gained popularity as it depends on the fact that speech signals are sparse in the time-frequency representations [8]. It tries to exploit some prior information of speech to recover speech components. The idea of sparsity suits audio signal [9], as speech exhibit a structured behavior like it is made of finite numbers of phonemes, formants, etc. The notion of sparsity for signal recovery has come from Compressed Sensing (CS) [10]. In general, noise signals are not structured, i.e., non-sparse, which has motivated authors to use speech sparsity for enhancement [11], [12]. In recent research, many authors have utilized sparsity to reduce the complexity of Machine Learning (ML) models [13], [14] as complexity is proportional to input feature size and has achieved improved performance. Thus sparsity- based techniques have massive scope in the future in the field of machine learning. In this paper, we will overview sparsity-based enhancement techniques and compare them to see which algorithm performs better under various noisy conditions.

## II. SPARSITY-BASED SPEECH ENHANCEMENT

A noisy speech signal $y$ is a mixture of clean signal $s$ and an additive noise $n$ as shown in (1).

$$y = s + n \tag{1}$$

All signals are real-valued one-dimensional signals. The representation of (1) in the transform domain is shown in (2).

$$Y = S + N \tag{2}$$

The transform can be Discrete Wavelet Transform (DWT) [16], Discrete Cosine Transform (DCT), Discrete Fourier transform (DFT), Real Discrete Gabor transform (RDGT) [15] etc. After transformation, the sparse representation of speech is given as $S = A\alpha$, where $A$ is the basis matrix of size L×M and $\alpha$ is a sparse vector of length $M$ with K << M number of the coefficients equal to zero. Now a speech enhancement can be referred to as recovering of sparse vector α for a given noisy signal spectrum Y. The enhancement problem based on speech sparsity can be formulated as:

$$\min \|\alpha\|_0 \text{ subject to } \|Y - A\alpha\|_2^2 < \epsilon \tag{3}$$

In equation (3), $\|\cdot\|_0$ represents $L_0$-norm, which counts the number of non-zero elements while the quadratic term puts a constraint on reconstruction, and $\epsilon$ represents the noise level. We will refer (3) as an $L_0$-minimization problem. If matrix A satisfies the coherence property, then there is a high probability that $L_1$-minimization, as shown in (4), gives solutions exactly as $L_0$-minimization [17].

$$\min \|\alpha\|_1 \text{ subject to } \|Y - A\alpha\|_2^2 < \epsilon \tag{4}$$

In equation (4), $\|\cdot\|_1$ represents the L1- norm of a signal $x$, which has been defined in (5).

$$\|x\|_1 = \sum_i |x(i)| \tag{5}$$

The Lagrange form of (4) is shown in (6).

$$\min \lambda\|\alpha\|_1 + \|Y - A\alpha\|_2^2 \tag{6}$$

Where $\lambda$ denotes the level of sparsity in the solution. The $\lambda$ should vary according to the signal's profile, e.g., the Gini index has been used in [18] for $\lambda$ estimation. The $L_1$-minimization is referred to as the Basis Pursuit Denoising (BPD) problem.

The basic structure of sparsity-based Speech Enhancement is shown in Fig. 1. The speech is segmented in the time domain using a short window with a duration of 20 to 40

978-1-6654-9099-3/23/$31.00 ©2023 IEEE

milliseconds. Afterward, the transform converts time-domain speech to a sparsely representable signal. The sparse vector α is recovered using solving the optimization problem shown in (3) or (4). The clean speech estimate in the transform domain is obtained by multiplication of $A$ and $\alpha$. Finally, time domain speech is obtained by inverse transform followed by overlap and add technique.

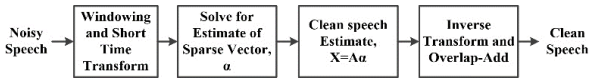


Fig. 1. Typical architecture of sparsity-based Speech Enhancement

The greedy approach is an intelligent way of searching for sparse solutions using a correlation between columns of the matrix A and measurement y. The major advantages of the greedy approach are its simple design and low computational time. The Orthogonal Matching Pursuit (OMP) [19] searches one column at a time, while the Compressive Sampling Matching Pursuit (CoSaMP) [20], [21], selects multiple columns in an iteration. CoSaMP has the capability to prune wrongly selected columns that OMP does not. The Matching Pursuit algorithms use the maximum inner product value between measurement y and columns of the matrix $A$ [21], [26]. In [27], authors have used The Least Angle Regression (LAR) algorithm, while in [28], authors have used mutual coherence for the search of columns.

In the Thresholding operation, all columns of $A$ are considered. In the simplest form, the coefficient of the least-square solution is retained above a particular threshold, and the rest is made zero in a single iteration. Iterative thresholding has been found to provide a better solution with constrained deviation from measurement $Y$. In [22], authors have used iterative thresholding for noise estimation by exploiting the behavior of noise-only and mixture frame to thresholding.

### A. Dictionary learning

In this approach, a dictionary whose columns represent various speech bases, known as atoms [23], is used. Spectrograms of speech signals were used to learn an over-complete dictionary by the K-singular value decomposition (K-SVD) algorithm [24]. In general, the magnitude of the spectrogram is used for dictionary learning, but power spectral density [25] can also be used as speech features. This learned dictionary better represents perceptual components of speech responsible for the naturalness than the transform matrix discussed earlier. This approach assumes that clean speech spectra $X$ can be represented using the linear combination of very few atoms of an overcomplete dictionary $D$ having $M$ atoms, as shown in Fig. 2. The optimization problem remains the same as (3) with basis matrix $A$ as $D$.

### B. Joint Dictionary Learning

In Joint Dictionary Learning (JDL) [25], [29], both speech and noise dictionaries are learned via alternatively conducting the sparse coding and dictionary update. It then concatenates these two dictionaries to provide one dictionary to recover clean speech from a noisy speech by solving a sparse optimization problem. This approach requires appropriate training data for noise to avoid voice activity detection. JDL algorithm is shown in Fig. 3. In it, dictionary $D = [D_x; D_d] \in R^{K\times(Lx+Ld)}$, is learned where $D_x$ is for clean speech, and $D_d$ is for noise. During the enhancement step, sparse vector of speech $\alpha_x \in R^{Lx}$ and noise $\alpha_d \in R^{Ld}$ are estimated by minimizing the approximation error [27] given by (3) with $A = D$. where $\alpha = [\alpha_x; \alpha_d]$.

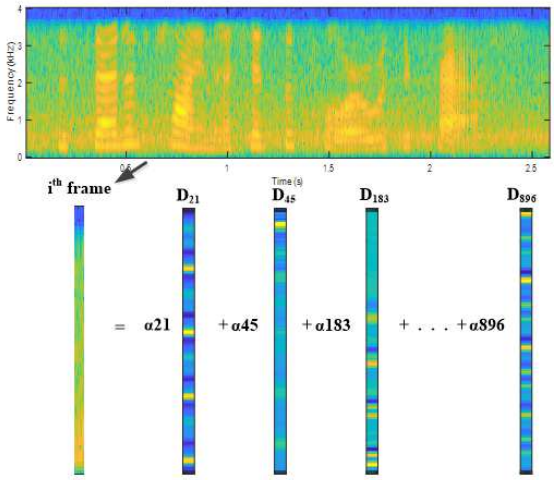


Fig. 2. Representation of a frame as a linear combination of atoms of the dictionary.

JDL still has a problem called source confusion, especially when noise shares atoms in $D_x$ and $D_d$, e.g., babble noise. In [32], the optimization technique minimizes the product of noise and speech dictionary, so that correlation between noise and speech dictionary is less and causes less confusion. Single JSR uses the mapping relationship between mixture and speech or between mixture and noise while Complementary Joint Sparse Representations (CJSR) [33], [34] uses both.

JDL-based methods have shown better performance for Ideal Binary Mask (IBM) estimation [36] compared to noise estimation-based techniques.

### C. Non-negative Matrix Factorization

In this approach, matrix $Y$ is formed by stacking several noisy speech-magnitude spectrum frames as columns in a time sequence. This approach uses sparsity of speech and low-rank structure of noise signal as noise spectra are generally highly correlated [37]. Non-negative Matrix Factorization(NMF) [38]–[40] decomposes the mixture data, $Y$ into speech, $S$ and noise, $N$ by solving the optimization problem shown in (7) using Robust Principle Component Analysis (RPCA) theory [41]. After getting $S$ and $N$, then gain is computed, and enhancement is achieved by multiplying gain with $Y$. For example, Wiener-type gain has been used in [39]

$$\min_{L,S} rank(L) + \lambda\|S\|_1 \text{ subject to } \|Y - S - L\|_2^2 < \epsilon \quad (7)$$

where λ is a positive parameter that balances the sparse term and the low-rank term. Generally, a low-rank solution of N is obtained using the thresholding of singular value decomposition (SVD) values.

This approach also does not require voice activity detection, similar to JDL. It performs well in low SNR conditions without knowing the distribution of noise signals. In the above problem, the rank and sparsity constraints are assumed to be known; hence, it is a supervised method.

Extracted speech matrix S, using RPCA theory, is too sparse and non-smooth, which causes loss of useful details

[42]. The decomposition of noisy speech to speech and background noise is less accurate in highly non-stationary, having a significant variation in the local SNR [45].

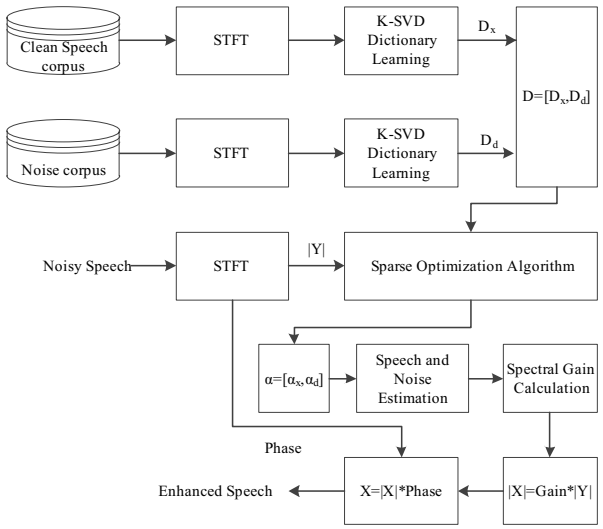


Fig. 3. Joint Dictionary Learning

## III. EXPERIMENT

In the experiment, the TIMIT database was used, having 40 sentences spoken by five male and five female speakers. All speech samples are down-sampled at 8 kHz. Noises used in the experiment are babble, car, F-16 cockpit, and factory environment. Noises have been added to clean speech with SNR from -10 dB to 5 dB. The Hamming window has been used for windowing with 50% overlapping in all experiments. All SSE algorithm uses a time-frequency domain for analysis of frame size 20ms. To assess the performance of enhancement techniques, Perceptual Evaluation of Speech Quality (PESQ) [46] has been used to measure the quality of the enhanced speech, and Short-Time Objective Intelligibility (STOI) [47] has been used to measure speech intelligibility. Both methods compare clean and enhanced speech temporal envelopes in several frames and bands. PESQ score lies between -0.5 to 4.5, and STOI lies between 0 and 1. Higher score values indicate better performance.

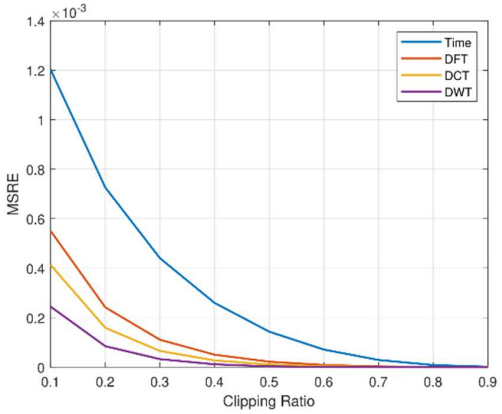


Fig. 4. Reconstruction error for various transform domains at different clipping ratios for a sentence

### A. Effect of various transform on reconstruction

As discussed in section 2, there are many transforms that help to represent speech signals sparsely in that domain. A signal $X = A\alpha$ is said to be sparse if most of the quantity of vector $\alpha$ is zero or nearly zero so that ignorance of the lower coefficient does not cause much degradation in the reconstructed signal. Here $A$ represents the transform basis. Let $x_T$ represent reconstructed after considering first $K$ ($K \ll M$) highest coefficients of vector $\alpha$, sorted in descending order.

The signal x will be sparse if Mean Squared Reconstruction Error (MSRE) given by (8) is zero or nearly zero. MSRE indicates the compressibility of speech as well as noise.

Fig. 4 shows the sparsity or compressibility of a speech signal in various domains like time-domain, DFT, DCT domain, and DWT (Daubechies of type 4). In the plot, the x-axis represents the clipping ratio, while the y-axis represents MSRE. The MSRE is lesser in the case of DWT than in DCT and DFT; hence signal is most sparse in DWT. Sparsity-based enhancement techniques will perform better in a transform if speech is more sparse in that domain compared to other [11].

$$MSRE = \frac{\sum_{i=1}^{N}(x(i)-x_T(i))^2}{N} \quad (8)$$

Fig. 5 shows the sparsity of various noises in the DCT domain, clearly showing that car noise is not as non-sparse as other noises. Hence the performance is not better in the case of car noises [21] as the inability of sparsity-based techniques to distinguish between noise and speech [11].

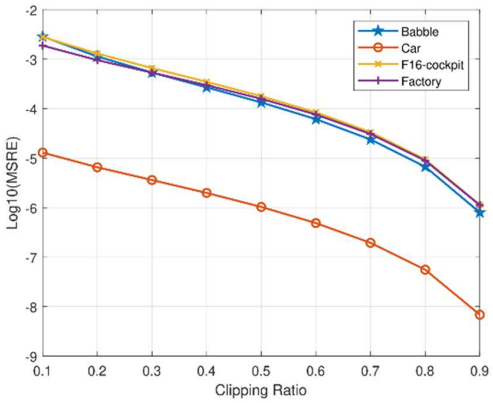


Fig. 5. Reconstruction error for various noises at different clipping ratios under the DCT domain.

TABLE I. LIST OF THE ENHANCEMENT TECHNIQUES

| Short form | Description | Reference |
|---|---|---|
| OMLSA | Optimally-Modified Log-Spectral Amplitude | [4] |
| CoSaMP | Compressive Sampling Matching Pursuit | [20], [21] |
| DL-SE | Dictionary Learning based Speech Enhancement | [23] |
| SS-OMP | Spectral Subtraction based OMP | [26] |
| JDL | Joint Dictionary Learning | [32] |
| NMF | Non-Negative Matrix Factorization | [7], [48] |

### B. Comparisons and discussions

For comparisons, several methods have been selected, as

shown in Table I. Out of the techniques, OMLSA is a conventional enhancement technique based on a statistical model. Fig. 6 and 7 show the PESQ and STOI scores of various techniques.

We will discuss the merits of sparsity-based enhancement techniques using results shown in Fig. 6 and 7.

1) Sparsity-based SE (SSE) requires some kind of speech presence estimation so that sparse motivated speech can be recovered from noisy speech; otherwise, it fails to differentiate between speech and noise region, especially at very low SNR [49]. To counter this issue, VAD is used to identify the speech region of the spectrogram, and only that region is enhanced. Other regions are masked in DL-SE technique [23], [26].
2) To retain speech components masked by the noise, both clean speech and noise dictionary knowledge are required. In JDL [32] and NMF [7], [48] technique, a noise dictionary is learned for each noise type. Therefore, the performance of these techniques is better compared to other techniques, especially the STOI score (Fig. 7).
3) Spectral Smoothing [21] of the spectrogram at low SNR significantly improves speech quality achieved by the CoSaMP technique.
4) Incorporating conventional noise estimation: In SS-OMP [26] technique, noise is subtracted before performing the CS reconstruction to reduce noise. Hence better PESQ score is achieved compared to other techniques, as shown in Fig. 6. The noise measurement can be estimated during pauses estimated using VAD.

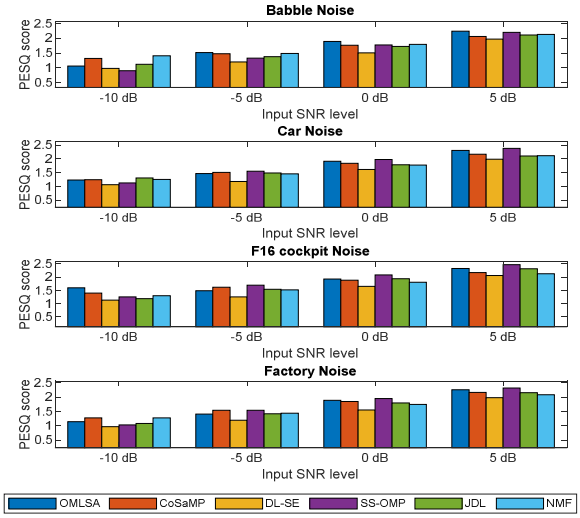


Fig. 6. PESQ performance of enhancement techniques

There are a few demerits that need to overlook.

1) The performance of CS improves as frame size increases [11], but this imposes considerable algorithmic delay, which makes it unfit for real-time application. In NMF [7], [48], a whole spectrogram of a speech of 2 to 4 seconds has been used.
2) All sparsity-based methods are iterative algorithms, not like the gain function in conventional methods, contributing to the computational delay.
3) Dictionary-based methods require a large amount of data for training, which may not be available in real-world scenarios. Almost all dictionary learning methods are offline, once learned, and do not adapt to incoming data that impose a lack of adaptability and may perform poorly as demography, dialect, and language change.

Overall, dictionary-based speech enhancement aided with voice activity detection and spectral subtraction performs better than others. The performances improve further if noise type and noise dictionary is known as apriori in the joint dictionary and non-negative matrix factorization method.

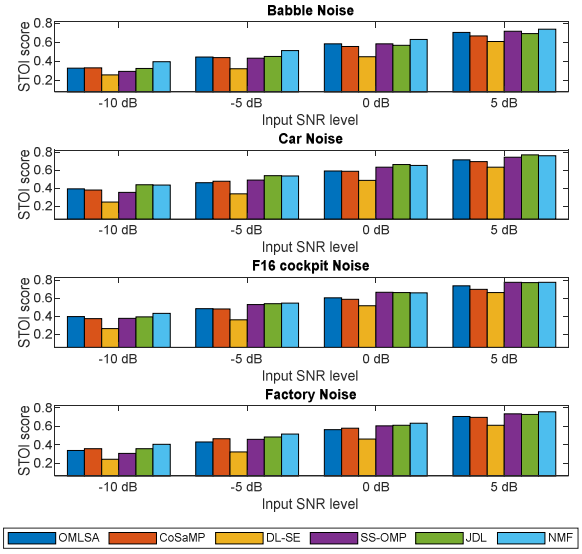


Fig. 7. STOI performance of enhancement techniques

## IV. Conclusion

This paper presents a comparative analysis of speech enhancement algorithms based on speech sparsity features. Several enhancement techniques have been investigated based on transform, dictionary learning, and the non-negative matrix factorization method, which promotes a sparse solution.

We compared several sparsity-based enhancement techniques in terms of PESQ and STOI scores on a common database. This analysis will help the researcher understand sparsity-based enhancement techniques and develop techniques to overcome existing issues. In future work, there is a need to look for such a transform domain where speech is highly sparse to accelerate the enhancement time as it depends on the sparsity of speech. The success of most of the sparsity-based approaches requires voice activity detection frame to frame and masking of noise- only components within the frame, which also needs to improve, especially under highly nonstationary noise, with the help of advanced machine learning models.

## References

[1] P. C. Loizou, Speech Enhancement Theory and Practice, 2nd ed. CRC Press, 2013.

[2] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," Speech Communication, vol. 49, no. 7-8, pp. 588–601, 2007.

[3] K. Paliwal, K. Wo´jcicki, and B. Schwerin, "Single-channel speech enhancement using spectral subtraction in the short-time modulation domain," Speech Communication, vol. 52, no. 5, pp. 450–475, may 2010.

[4] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," IEEE Transactions on Speech and Audio Processing, vol. 11, no. 5, pp. 466–475, 2003.

[5] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 4, pp. 1383–1393, may 2012.

[6] Y. Ephraim and D. Malah, "Speech enhancement using a- Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," IEEE Transactions on Audio, Speech and Language Processing, vol. ASSP-32, no. 6, pp. 1109–1121, 1984.

[7] H. Chung, E. Plourde, and B. Champagne, "Regularized non-negative matrix factorization with Gaussian mixtures and masking model for speech enhancement," Speech Communication, vol. 87, pp. 18–30, 2017.

[8] T. J. Gardner and M. O. Magnasco, "Sparse time-frequency representations," Proceedings of the National Academy of Sciences, vol. 103, no. 16, pp. 6094 LP – 6099, apr 2006.

[9] M. G. Christensen, J. Ostergaard, and S. H. Jensen, "On compressed sensing and its application to speech and audio signals," in 2009 Conference Record of the Forty-Third Asilomar Conference on Signals, Systems and Computers. IEEE, 2009, pp. 356–360.

[10] R. G. Baraniuk, E. Candes, M. Elad, and Y. Ma, "Applications of sparse representation and compressive sensing," Proceedings of the IEEE, vol. 98, no. 6, pp. 906–909, 2010.

[11] S. Y. Low, D. S. Pham, and S. Venkatesh, "Compressive speech enhancement," Speech Communication, vol. 55, no. 6, pp. 757–768, 2013.

[12] D. Wu, W. P. Zhu, and M. N. Swamy, "A compressive sensing method for noise reduction of speech and audio signals," Midwest Symposium on Circuits and Systems, pp. 0–3, 2011.

[13] M. I. Khattak, N. Saleem, J. Gao, E. Verdu, and J. P. Fuente, "Regularized sparse features for noisy speech enhancement using deep neural networks," Computers and Electrical Engineering, vol. 100, no. August 2021, p. 107887, 2022.

[14] A. Garg, "Speech enhancement using long short term memory with trained speech features and adaptive wiener filter," Multimedia Tools and Applications, 2022.

[15] J. Wen, Z. Chu, and J. Zhou, "Evaluation of intelligibility of noisy whisper enhanced by compressive sensing," Proceedings of 2nd Interna- tional Conference on Information Technology and Electronic Commerce, ICITEC 2014, pp. 216–219, 2014.

[16] S. Sahu and N. Rayavarapu, "Performance comparison of sparsifying basis functions for compressive speech enhancement," International Journal of Speech Technology, vol. 22, no. 3, pp. 769–783, 2019.

[17] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries introduction," IEEE Transactions on Image Processing, vol. 15, no. 12, pp. 3736– 3745, 2011.

[18] Y. Shi, S. Y. Low, and K. F. Cedric Yiu, "Hyper-parameterization of sparse reconstruction for speech enhancement," Applied Acoustics, vol. 138, no. September 2017, pp. 72–79, 2018.

[19] H. Yang, D. Hao, H. Sun, and Y. Liu, "Speech enhancement using orthogonal matching pursuit algorithm," IEEE International Conference on Orange Technologies, ICOT 2014, pp. 101–104, 2014.

[20] D. Wu, W. P. Zhu, and M. N. Swamy, "Compressive sensing-based speech enhancement in non-sparse noisy environments," IET Signal Processing, vol. 7, no. 5, pp. 450–457, 2013.

[21] ——, "The theory of compressive sensing matching pursuit considering time-domain noise with application to speech enhancement," IEEE Transactions on Audio, Speech and Language Processing, vol. 22, no. 3, pp. 682–696, 2014.

[22] R. Kumar, M. Tripathy, and R. S. Anand, "Iterative thresholding-based spectral subtraction algorithm for speech enhancement," in Advances in VLSI, Signal Processing, Power Electronics, IoT, Communication and Embedded Systems. Springer, 2021, ch. 18, pp. 221–232.

[23] J. C. Wang, Y. S. Lee, C. H. Lin, S. F. Wang, C. H. Shih, and C. H. Wu, "Compressive sensing-based speech enhancement," IEEE/ACM Transactions on Audio Speech and Language Processing, vol. 24, no. 11, pp. 2122–2131, 2016.

[24] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," IEEE Transactions on Signal Processing, vol. 54, no. 11, pp. 4311–4322, 2006.

[25] Y. Zhao, X. Zhao, and B. Wang, "A speech enhancement method employing sparse representation of power spectral density," Journal of Information and Computational Science, vol. 10, no. 6, pp. 1705–1714, 2013.

[26] H. Haneche, B. Boudraa, and A. Ouahabi, "A new way to enhance speech signal based on compressed sensing," Measurement, vol. 151, feb 2020.

[27] C. D. Sigg, T. Dikk, and 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.

[28] S. Mavaddaty, S. M. Ahadi, and S. Seyedin, "Modified coherence-based dictionary learning method for speech enhancement," IET Signal Processing, vol. 9, no. 7, pp. 537–545, 2015.

[29] C. D. Sigg, T. Dikk, and J. M. Buhmann, "Speech enhancement with sparse coding in learned dictionaries," ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, pp. 4758–4761, 2010.

[30] Y. He, J. Han, S. Deng, T. Zheng, and G. Zheng, "A solution to residual noise in speech denoising with sparse representation," ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, pp. 4653–4656, 2012.

[31] W. Li, Y. Zhou, N. Poh, F. Zhou, and Q. Liao, "Feature denoising using joint sparse representation for in-car speech recognition," IEEE Signal Processing Letters, vol. 20, no. 7, pp. 681–684, 2013.

[32] L. Sun, Y. Bu, P. Li, and Z. Wu, "Single-channel speech enhancement based on joint constrained dictionary learning," EURASIP Journal on Audio, Speech, and Music Processing, vol. 2021, no. 1, pp. 1–14, dec 2021.

[33] Y. Luo, G. Bao, Y. Xu, and Z. Ye, "Supervised monaural speech enhancement using complementary joint sparse representations," IEEE SIGNAL PROCESSING LETTERS, vol. 23, no. 2, pp. 237–241, 2016.

[34] J. Fu, L. Zhang, and Z. Ye, "Supervised monaural speech enhancement using two-level complementary joint sparse representations," Applied Acoustics, vol. 132, no. October 2017, pp. 1–7, 2018.

[35] L. Huang, L. Li, and S. He, "Speech enhancement based on sparse representation using universal dictionary," in Proceedings of the International Conference on Anti-Counterfeiting, Security and Identification, ASID. IEEE, 2013, pp. 9–12.

[36] J. Sun, Y. Tang, A. Jiang, N. Xu, and L. Zhou, "Speech enhancement via sparse coding with ideal binary mask," International Conference on Signal Processing Proceedings, ICSP, vol. 2015-Janua, no. October, pp. 537–540, 2014.

[37] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," IEEE Transactions on Audio, Speech and Language Processing, vol. 15, no. 3, pp. 1066–1074, 2007.

[38] H. T. Fan, J. W. Hung, X. Lu, S. S. Wang, and Y. Tsao, "Speech enhancement using segmental nonnegative matrix factorization," ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, pp. 4483–4487, 2014.

[39] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," IEEE Transactions on Audio, Speech and Language Processing, vol. 21, no. 10, pp. 2140–2151, 2013.

[40] M. Sun, Y. Li, J. F. Gemmeke, and X. Zhang, "Speech enhancement under low SNR conditions via noise estimation using sparse and low-rank NMF with Kullback-Leibler divergence," IEEE Transactions on Audio, Speech and Language Processing, vol. 23, no. 7, pp. 1233–1242, 2015.

[41] E. J. Cand, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis ?" Journal of the ACM, vol. 58, no. 3, pp. 1–39, 2011.

[42] Y. Li, X. Zhang, M. Sun, and G. Min, "Unsupervised monaural speech enhancement using robust NMF with low-rank and sparse constraints," in IEEE China Summit and International Conference on Signal and Information Processing. Chengdu: IEEE, 2015, pp. 1–4.

[43] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 1, pp. 1–12, 2007.

[44] M. A. Carlin, N. Malyska, and T. F. Quatieri, "Speech enhancement using sparse convolutive non-negative matrix factorization with basis adaptation," 13th Annual Conference of the International Speech Com-

munication Association 2012, INTERSPEECH 2012, vol. 1, pp. 582–585, 2012.

[45] Z. Chen, B. McFee, and D. P. Ellis, "Speech enhancement by low-rank and convolutive dictionary spectrogram decomposition," in Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, no. 1, 2014, pp. 2833–2837.

[46] "Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," ITU, ITU-T Recommendation P. 862, 2000.

[47] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short- time objective intelligibility measure for time-frequency weighted noisy speech," IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 4214–4217, 2010.

[48] Y. Xiang, L. Shi, J. L. Hojvang, M. Hojfeldt Rasmussen, and M. G. Christensen, "A Novel NMF-HMM speech enhancement algorithm based on poisson mixture model," in ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, jun 2021, pp. 721–725.

[49] L. Wang, D. Wang, and C. Hao, "A multiple-measurement vectors reconstruction method for low SNR scenarios," IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 67, no. 4, pp. 785–789, 2020