REFERENCES

[1] Abdulatif, S., Cao, R., & Yang, B. (2024). CMGAN: Conformer-Based Metric-GAN for Monaural Speech Enhancement. *IEEE/ACM Transactions on Audio Speech and Language Processing*, *32*, 2477–2493. https://doi.org/10.1109/TASLP.2024.3393718

[2] Aguilar-Torres, G., Nakano-Miyatake, M., & Perez-Meana, H. (2006). Enhancement and Restoration of Alaryngeal Speech Signals. *16th International Conference on Electronics, Communications and Computers (CONIELECOMP'06)*, 31–31. https://doi.org/10.1109/CONIELECOMP.2006.29

[3] Ahmadi, F., Kobayashi, K., & Toda, T. (2019). Development of a Real-time Bionic Voice Generation System based on Statistical Excitation Prediction. *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, 655–657. https://doi.org/10.1145/3308561.3354591

[4] Allen, J. B., & Rabiner, L. R. (1977). A unified approach to short-time Fourier analysis and synthesis. *Proceedings of the IEEE*, *65*(11), 1558–1564. https://doi.org/10.1109/PROC.1977.10770

[5] Azarang, A., & Kehtarnavaz, N. (2020). A review of multi-objective deep learning speech denoising methods. *Speech Communication*, *122*, 1–10. https://doi.org/10.1016/j.specom.2020.04.002

[6] Baby, D., Virtanen, T., Gemmeke, J. F., & Van hamme, H. (2015). Coupled Dictionaries for Exemplar-Based Speech Enhancement and Automatic Speech Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *23*(11), 1788–1799. https://doi.org/10.1109/TASLP.2015.2450491

[7] Bao, F., & Abdulla, W. H. (2019). A New Ratio Mask Representation for CASA-Based Speech Enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *27*(1), 7–19. https://doi.org/10.1109/TASLP.2018.2868407

[8] Bishop, C. M. (2007). Pattern Recognition and Machine Learning. *Journal of Electronic Imaging*, *16*(4), 049901. https://doi.org/10.1117/1.2819119

[9] Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *27*(2), 113–120. https://doi.org/10.1109/TASSP.1979.1163209

[10] Brook, I., & Goodman, J. F. (2020). Tracheoesophageal Voice Prosthesis Use and Maintenance in Laryngectomees. *International Archives of Otorhinolaryngology*, *24*(04), e535–e538. https://doi.org/10.1055/s-0039-3402497

[11] Chen, J., & Liang, Z. (2018). The Application of Deep Neural Network in Speech Enhancement Processing. *2018 5th International Conference on Information Science and Control Engineering (ICISCE)*, 1263–1266. https://doi.org/10.1109/ICISCE.2018.00257

[12] Cheng, R., Bao, C., & Xiang, Y. (2018). Speech Enhancement with Phase Correction based on Modified DNN Architecture. *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 1222–1227. https://doi.org/10.23919/APSIPA.2018.8659625

[13] Chern, A., Lai, Y.-H., Chang, Y.-P., Tsao, Y., Chang, R. Y., & Chang, H.-W. (2017). A Smartphone-Based Multi-Functional Hearing Assistive System to Facilitate Speech Recognition in the Classroom. *IEEE Access*, *5*, 10339–10351. https://doi.org/10.1109/ACCESS.2017.2711489

[14] Chung, H., Kim, T., Plourde, E., & Champagne, B. (2018). NOISE-ADAPTIVE DEEP NEURAL NETWORK FOR SINGLE-CHANNEL SPEECH ENHANCEMENT. *2018 IEEE*

28th International Workshop on Machine Learning for Signal Processing (MLSP), 1–6. https://doi.org/10.1109/MLSP.2018.8517027

[15] Cohen, I. (2003). Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging. *IEEE Transactions on Speech and Audio Processing*, *11*(5), 466–475. https://doi.org/10.1109/TSA.2003.811544

[16] Das, A., & Hansen, J. H. L. (2012). Constrained Iterative Speech Enhancement Using Phonetic Classes. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(6), 1869–1883. https://doi.org/10.1109/TASL.2012.2191282

[17] Dietz, A. (2004). Epidemiology of Laryngeal Cancer. *Laryngo-Rhino-Otologie*, *83*(11), 771–772. https://doi.org/10.1055/s-2004-826025

[18] Dionelis, N., & Brookes, M. (2018). Phase-Aware Single-Channel Speech Enhancement With Modulation-Domain Kalman Filtering. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(5), 937–950. https://doi.org/10.1109/TASLP.2018.2800525

[19] Doi, H., Toda, T., Nakamura, K., Saruwatari, H., & Shikano, K. (2014). Alaryngeal Speech Enhancement Based on One-to-Many Eigenvoice Conversion. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *22*(1), 172–183. https://doi.org/10.1109/TASLP.2013.2286917

[20] Dubey, S. R., Singh, S. K., & Chaudhuri, B. B. (2022). Activation functions in deep learning: A comprehensive survey and benchmark. *Neurocomputing*, *503*, 92–108. https://doi.org/10.1016/j.neucom.2022.06.111

[21] Ephraim, Y., & Malah, D. (1985). Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *33*(2), 443–445. https://doi.org/10.1109/TASSP.1985.1164550

[22] Fu, S.-W., Wang, T.-W., Tsao, Y., Lu, X., & Kawai, H. (2018). End-to-End Waveform Utterance Enhancement for Direct Evaluation Metrics Optimization by Fully Convolutional Neural Networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(9), 1570–1584. https://doi.org/10.1109/TASLP.2018.2821903

[23] Gelderblom, F. B., Tronstad, T. V., & Viggen, E. M. (2019). Subjective Evaluation of a Noise-Reduced Training Target for Deep Neural Network-Based Speech Enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *27*(3), 583–594. https://doi.org/10.1109/TASLP.2018.2882738

[24] Gholamalinezhad, H., & Khosravi, H. (2020). *Pooling Methods in Deep Neural Networks, a Review*.

[25] Glorot, X., Bordes, A., & Bengio, Y. (2011). *Deep Sparse Rectifier Neural Networks*.

[26] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*.

[27] Gowda, D., Kadiri, S. R., Story, B., & Alku, P. (2020). Time-varying quasi-closed-phase analysis for accurate formant tracking in speech signals. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 1–1. https://doi.org/10.1109/TASLP.2020.3000037

[28] Grais, E. M., Roma, G., Simpson, A. J. R., & Plumbley, M. D. (2017a). *Discriminative Enhancement for Single Channel Audio Source Separation Using Deep Neural Networks* (pp. 236–246). https://doi.org/10.1007/978-3-319-53547-0_23

[29] Grais, E. M., Roma, G., Simpson, A. J. R., & Plumbley, M. D. (2017b). Two-Stage Single-Channel Audio Source Separation Using Deep Neural Networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(9), 1773–1783. https://doi.org/10.1109/TASLP.2017.2716443

[30] Grzywalski, T., & Drgas, S. (2022). Speech Enhancement by Multiple Propagation through the Same Neural Network. *Sensors*, *22*(7), 2440. https://doi.org/10.3390/s22072440

[31] Gupta, T. K., & Raza, K. (2020). Optimizing Deep Feedforward Neural Network Architecture: A Tabu Search Based Approach. *Neural Processing Letters*, *51*(3), 2855–2870. https://doi.org/10.1007/s11063-020-10234-7

[32] Hasannezhad, M., Ouyang, Z., Zhu, W.-P., & Champagne, B. (2021). Speech Enhancement With Phase Sensitive Mask Estimation Using a Novel Hybrid Neural Network. *IEEE Open Journal of Signal Processing*, *2*, 136–150. https://doi.org/10.1109/OJSP.2021.3067147

[33] He, Q., Bao, F., & Bao, C. (2017). Multiplicative Update of Auto-Regressive Gains for Codebook-Based Speech Enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(3), 457–468. https://doi.org/10.1109/TASLP.2016.2636445

[34] Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *Neural and Evolutionary Computing*.

[35] Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[36] Hoffmann, T. K. (2021). Total Laryngectomy—Still Cutting-Edge? *Cancers*, *13*(6), 1405. https://doi.org/10.3390/cancers13061405

[37] Honda, K. (2008). Physiological Processes of Speech Production. In *Springer Handbook of Speech Processing* (pp. 7–26). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-49127-9_2

[38] Hsieh, T.-A., Wang, H.-M., Lu, X., & Tsao, Y. (2020). WaveCRN: An Efficient Convolutional Recurrent Neural Network for End-to-End Speech Enhancement. *IEEE Signal Processing Letters*, *27*, 2149–2153. https://doi.org/10.1109/LSP.2020.3040693

[39] Huang, Q., Bao, C., Wang, X., & Xiang, Y. (2018). DNN-Based Speech Enhancement Using MBE Model. *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 196–200. https://doi.org/10.1109/IWAENC.2018.8521278

[40] IEEE Recommended Practice for Speech Quality Measurements. (1969). *IEEE Transactions on Audio and Electroacoustics*, *17*(3), 225–246. https://doi.org/10.1109/TAU.1969.1162058

[41] Ioffe, S., & Szegedy, C. (2015). *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*.

[42] Jaiswal, R. K., Yeduri, S. R., & Cenkeramaddi, L. R. (2022). Single-channel speech enhancement using implicit Wiener filter for high-quality speech communication. *International Journal of Speech Technology*, *25*(3), 745–758. https://doi.org/10.1007/s10772-022-09987-4

[43] Kameoka, H., Tanaka, K., Kwasny, D., Kaneko, T., & Hojo, N. (2020). ConvS2S-VC: Fully Convolutional Sequence-to-Sequence Voice Conversion. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, 1849–1863. https://doi.org/10.1109/TASLP.2020.3001456

[44] Kang, T. G., Kwon, K., Shin, J. W., & Kim, N. S. (2015). NMF-based Target Source Separation Using Deep Neural Network. *IEEE Signal Processing Letters*, *22*(2), 229–233. https://doi.org/10.1109/LSP.2014.2354456

[45] Kawase, T., Okamoto, M., Fukutomi, T., & Takahashi, Y. (2020). Speech Enhancement Parameter Adjustment to Maximize Accuracy of Automatic Speech Recognition. *IEEE Transactions on Consumer Electronics*, *66*(2), 125–133. https://doi.org/10.1109/TCE.2020.2986003

[46] Kaye, R., Tang, C. G., & Sinclair, C. F. (2017). The electrolarynx: voice restoration after total laryngectomy. *Medical Devices: Evidence and Research*, *Volume 10*, 133–140. https://doi.org/10.2147/MDER.S133225

[47] Kim, G., Lee, H., Kim, B.-K., Oh, S.-H., & Lee, S.-Y. (2019). Unpaired Speech Enhancement by Acoustic and Adversarial Supervision for Speech Recognition. *IEEE Signal Processing Letters*, *26*(1), 159–163. https://doi.org/10.1109/LSP.2018.2880285

[48] Kramp, B., & Dommerich, S. (2009). *Tracheostomy cannulas and voice prosthesis*. https://doi.org/10.3205/cto000057

[49] Krawczyk, M., & Gerkmann, T. (2014). STFT Phase Reconstruction in Voiced Speech for an Improved Single-Channel Speech Enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *22*(12), 1931–1940. https://doi.org/10.1109/TASLP.2014.2354236

[50] Kumar Shukla, N., Shajin, F. H., & Rajendran, R. (2024). Speech enhancement system using deep neural network optimized with Battle Royale Optimization. *Biomedical Signal Processing and Control*, *92*, 105991. https://doi.org/10.1016/j.bspc.2024.105991

[51] Lan, T., Lyu, Y., Ye, W., Hui, G., Xu, Z., & Liu, Q. (2020). Combining Multi-Perspective Attention Mechanism With Convolutional Networks for Monaural Speech Enhancement. *IEEE Access*, *8*, 78979–78991. https://doi.org/10.1109/ACCESS.2020.2989861

[52] Lavanya, T., Nagarajan, T., & Vijayalakshmi, P. (2020). Multi-Level Single-Channel Speech Enhancement Using a Unified Framework for Estimating Magnitude and Phase Spectra. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, 1315–1327. https://doi.org/10.1109/TASLP.2020.2986877

[53] LeCun, Y. A., Bottou, L., Orr, G. B., & Müller, K.-R. (2012). *Efficient BackProp* (pp. 9–48). https://doi.org/10.1007/978-3-642-35289-8_3

[54] Lee, H.-Y., Cho, J.-W., Kim, M., & Park, H.-M. (2016). DNN-Based Feature Enhancement Using DOA-Constrained ICA for Robust Speech Recognition. *IEEE Signal Processing Letters*, *23*(8), 1091–1095. https://doi.org/10.1109/LSP.2016.2583658

[55] Lee, J.-W., Kim, S., & Kang, H.-G. (2014). Detecting pathological speech using contour modeling of harmonic-to-noise ratio. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5969–5973. https://doi.org/10.1109/ICASSP.2014.6854749

[56] Levitt, H. (2001). Noise reduction in hearing aids: a review. *Journal of Rehabilitation Research and Development*, *38*(1), 111–121.

[57] Li, A., Zheng, C., Yu, G., Cai, J., & Li, X. (2022). Filtering and Refining: A Collaborative-Style Framework for Single-Channel Speech Enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *30*, 2156–2172. https://doi.org/10.1109/TASLP.2022.3184889

[58] Li, B., Tsao, Y., & Sim, K. C. (2013). An investigation of spectral restoration algorithms for deep neural networks based noise robust speech recognition. *Interspeech 2013*, 3002–3006. https://doi.org/10.21437/Interspeech.2013-278

[59] Li, J., Yang, L., Zhang, J., Yan, Y., Hu, Y., Akagi, M., & Loizou, P. C. (2011). Comparative intelligibility investigation of single-channel noise-reduction algorithms for Chinese, Japanese, and English. *The Journal of the Acoustical Society of America*, *129*(5), 3291–3301. https://doi.org/10.1121/1.3571422

[60] Li, R., Liu, Y., Shi, Y., Dong, L., & Cui, W. (2016). ILMSAF based speech enhancement with DNN and noise classification. *Speech Communication*, *85*, 53–70. https://doi.org/10.1016/j.specom.2016.10.008

[61] Li, R., Sun, X., Li, T., & Zhao, F. (2020). A multi-objective learning speech enhancement algorithm based on IRM post-processing with joint estimation of SCNN and TCNN. *Digital Signal Processing*, *101*, 102731. https://doi.org/10.1016/j.dsp.2020.102731

[62] Liang, R., Kong, F., Xie, Y., Tang, G., & Cheng, J. (2020). Real-Time Speech Enhancement Algorithm Based on Attention LSTM. *IEEE Access*, *8*, 48464–48476. https://doi.org/10.1109/ACCESS.2020.2979554

[63] Lim, J. S., & Oppenheim, A. V. (1979). Enhancement and bandwidth compression of noisy speech. *Proceedings of the IEEE*, *67*(12), 1586–1604. https://doi.org/10.1109/PROC.1979.11540

[64] Liu, Q., Wang, W., Jackson, P. J. B., & Tang, Y. (2017). A perceptually-weighted deep neural network for monaural speech enhancement in various background noise conditions. *2017 25th European Signal Processing Conference (EUSIPCO)*, 1270–1274. https://doi.org/10.23919/EUSIPCO.2017.8081412

[65] Lu, Y., & Salem, F. M. (2017). *Simplified Gating in Long Short-term Memory (LSTM) Recurrent Neural Networks*.

[66] Ma, Y., & Nishihara, A. (2014). A modified Wiener filtering method combined with wavelet thresholding multitaper spectrum for speech enhancement. *EURASIP Journal on Audio, Speech, and Music Processing*, *2014*(1), 32. https://doi.org/10.1186/s13636-014-0032-7

[67] Ming, J., & Crookes, D. (2017). Speech Enhancement Based on Full-Sentence Correlation and Clean Speech Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(3), 531–543. https://doi.org/10.1109/TASLP.2017.2651406

[68] Mouchtaris, A., Van der Spiegel, J., & Mueller, P. (n.d.). A spectral conversion approach to the iterative Wiener filter for speech enhancement. *2004 IEEE International Conference on Multimedia and Expo (ICME) (IEEE Cat. No.04TH8763)*, 1971–1974. https://doi.org/10.1109/ICME.2004.1394648

[69] Mowlaee, P., Blass, M., & Kleijn, W. B. (2017). New Results in Modulation-Domain Single-Channel Speech Enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(11), 2125–2137. https://doi.org/10.1109/TASLP.2017.2747082

[70] Mowlaee, P., & Stahl, J. K. W. (2020). Single-channel speech enhancement with correlated spectral components: Limits-potential. *Speech Communication*, *121*, 58–69. https://doi.org/10.1016/j.specom.2020.05.002

[71] Narwaria, M., Lin, W., McLoughlin, I. V., Emmanuel, S., & Chia, L.-T. (2012). Nonintrusive Quality Assessment of Noise Suppressed Speech With Mel-Filtered Energies and Support Vector Regression. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(4), 1217–1232. https://doi.org/10.1109/TASL.2011.2174223

[72] Pandey, A., & Wang, D. (2019). A New Framework for CNN-Based Speech Enhancement in the Time Domain. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *27*(7), 1179–1188. https://doi.org/10.1109/TASLP.2019.2913512

[73] Pandey, A., & Wang, D. (2021). Dense CNN With Self-Attention for Time-Domain Speech Enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 1270–1279. https://doi.org/10.1109/TASLP.2021.3064421

[74] Phan, H., McLoughlin, I. V., Pham, L., Chen, O. Y., Koch, P., De Vos, M., & Mertins, A. (2020). Improving GANs for Speech Enhancement. *IEEE Signal Processing Letters*, *27*, 1700–1704. https://doi.org/10.1109/LSP.2020.3025020

[75] Ramachandran, P., Zoph, B., & Le, Q. V. (2017). *Searching for Activation Functions*.

[76] Rascon, C. (2023). Characterization of Deep Learning-Based Speech-Enhancement Techniques in Online Audio Processing Applications. *Sensors*, *23*(9), 4394. https://doi.org/10.3390/s23094394

[77] Re, D. E., O'Connor, J. J. M., Bennett, P. J., & Feinberg, D. R. (2012). Preferences for Very Low and Very High Voice Pitch in Humans. *PLoS ONE*, *7*(3), e32719. https://doi.org/10.1371/journal.pone.0032719

[78] Resch, B., Nilsson, M., Ekman, A., & Kleijn, W. B. (2007). Estimation of the Instantaneous Pitch of Speech. *IEEE Transactions on Audio, Speech and Language Processing*, *15*(3), 813–822. https://doi.org/10.1109/TASL.2006.885242

[79] Roy, S. K., Nicolson, A., & Paliwal, K. K. (2021). DeepLPC: A Deep Learning Approach to Augmented Kalman Filter-Based Single-Channel Speech Enhancement. *IEEE Access*, *9*, 64524–64538. https://doi.org/10.1109/ACCESS.2021.3075209

[80] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. https://doi.org/10.1038/323533a0

[81] Saleem, N., Irfan Khattak, M., Nawaz, A., Umer, F., & Kumar Ochani, M. (2021). Perceptually weighted β-order spectral amplitude Bayesian estimator for phase compensated speech enhancement. *Applied Acoustics*, *178*, 108007. https://doi.org/10.1016/j.apacoust.2021.108007

[82] Saleem, N., Irfan, M., Chen, X., & Ali, M. (2018). Deep Neural Network based Supervised Speech Enhancement in Speech-Babble Noise. *2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS)*, 871–874. https://doi.org/10.1109/ICIS.2018.8466542

[83] Saleem, N., Khattak, M. I., Al-Hasan, M., & Jan, A. (2021). Multi-objective long-short term memory recurrent neural networks for speech enhancement. *Journal of Ambient Intelligence and Humanized Computing*, *12*(10), 9037–9052. https://doi.org/10.1007/s12652-020-02598-4

[84] Saleem, N., Khattak, M. I., Al-Hasan, M., & Qazi, A. B. (2020). On Learning Spectral Masking for Single Channel Speech Enhancement Using Feedforward and Recurrent Neural Networks. *IEEE Access*, *8*, 160581–160595. https://doi.org/10.1109/ACCESS.2020.3021061

[85] Sarker, I. H. (2021). Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *SN Computer Science*, *2*(6), 420. https://doi.org/10.1007/s42979-021-00815-1

[86] Seltzer, M. L., Yu, D., & Wang, Y. (2013). An investigation of deep neural networks for noise robust speech recognition. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 7398–7402. https://doi.org/10.1109/ICASSP.2013.6639100

[87] Shabanian, S., Arpit, D., Trischler, A., & Bengio, Y. (2017). *Variational Bi-LSTMs*.

[88] Shahnaz, C., Zhu, W. -p., & Ahmad, M. O. (2006). A New Technique for the Estimation of Jitter and Shimmer of Voiced Speech Signal. *2006 Canadian Conference on Electrical and Computer Engineering*, 2112–2115. https://doi.org/10.1109/CCECE.2006.277799

[89] Sharifzadeh, H. R., McLoughlin, I. V., & Ahmadi, F. (2010). Reconstruction of Normal Sounding Speech for Laryngectomy Patients Through a Modified CELP Codec. *IEEE Transactions on Biomedical Engineering*, *57*(10), 2448–2458. https://doi.org/10.1109/TBME.2010.2053369

[90] Shimada, K., Bando, Y., Mimura, M., Itoyama, K., Yoshii, K., & Kawahara, T. (2019). Unsupervised Speech Enhancement Based on Multichannel NMF-Informed Beamforming for Noise-Robust Automatic Speech Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *27*(5), 960–971. https://doi.org/10.1109/TASLP.2019.2907015

[91] Širić, L., Rosso, M., & Včev, A. (2018). The Role of Esophagus in Voice Rehabilitation of Laryngectomees. In *Esophageal Cancer and Beyond*. InTech. https://doi.org/10.5772/intechopen.78594

[92] Sitzmann, V., Martel, J. N. P., Bergman, A. W., Lindell, D. B., & Wetzstein, G. (2020). *Implicit Neural Representations with Periodic Activation Functions*.

[93] Staudemeyer, R. C., & Morris, E. R. (2019). *Understanding LSTM -- a tutorial into Long Short-Term Memory Recurrent Neural Networks*.

[94] Strake, M., Defraene, B., Fluyt, K., Tirry, W., & Fingscheidt, T. (2020). Fully Convolutional Recurrent Networks for Speech Enhancement. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6674–6678. https://doi.org/10.1109/ICASSP40776.2020.9054230

[95] Sze, V., Chen, Y.-H., Yang, T.-J., & Emer, J. S. (2017). Efficient Processing of Deep Neural Networks: A Tutorial and Survey. *Proceedings of the IEEE*, *105*(12), 2295–2329. https://doi.org/10.1109/JPROC.2017.2761740

[96] Takeuchi, D., Yatabe, K., Koizumi, Y., Oikawa, Y., & Harada, N. (2020). Real-Time Speech Enhancement Using Equilibriated RNN. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 851–855. https://doi.org/10.1109/ICASSP40776.2020.9054597

[97] Tamilselvan, P., Yibin Wang, & Pingfeng Wang. (2012). Deep Belief Network based state classification for structural health diagnosis. *2012 IEEE Aerospace Conference*, 1–11. https://doi.org/10.1109/AERO.2012.6187366

[98] Tang, X., Du, J., Chai, L., Wang, Y., Wang, Q., & Lee, C.-H. (2019). A LSTM-Based Joint Progressive Learning Framework for Simultaneous Speech Dereverberation and Denoising. *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 274–278. https://doi.org/10.1109/APSIPAASC47483.2019.9023160

[99] Tantibundhit, C., Pernkopf, F., & Kubin, G. (2010). Joint Time–Frequency Segmentation Algorithm for Transient Speech Decomposition and Speech Enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*(6), 1417–1428. https://doi.org/10.1109/TASL.2009.2035037

[100] Tkachenko, M., Yamshinin, A., Lyubimov, N., Kotov, M., & Nastasenko, M. (2017). *Speech Enhancement for Speaker Recognition Using Deep Recurrent Neural Networks* (pp. 690–699). https://doi.org/10.1007/978-3-319-66429-3_69

[101] Tu, M., & Zhang, X. (2017). Speech enhancement based on Deep Neural Networks with skip connections. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5565–5569. https://doi.org/10.1109/ICASSP.2017.7953221

[102] Tu, Y.-H., Du, J., & Lee, C.-H. (2019). Speech Enhancement Based on Teacher–Student Deep Learning Using Improved Speech Presence Probability for Noise-Robust Speech Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *27*(12), 2080–2091. https://doi.org/10.1109/TASLP.2019.2940662

[103] Tu, Y.-H., Du, J., & Lee, C.-H. (2020). 2D-to-2D Mask Estimation for Speech Enhancement Based on Fully Convolutional Neural Network. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6664–6668. https://doi.org/10.1109/ICASSP40776.2020.9054615

[104] Tuzla, O. (2011). SINGLE CHANNEL SPEECH MUSIC SEPARATION USING NONNEGATIVE MATRIX FACTORIZATION AND SPECTRAL MASKS Emad M . Grais and Hakan Erdogan Faculty of Engineering and Natural Sciences ,. *Digital Signal Processing (DSP), 2011 …*. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6004924

[105] Tyburek, K. (2022). Parameterisation of human speech after total laryngectomy surgery. *Computer Speech & Language*, *72*, 101313. https://doi.org/10.1016/j.csl.2021.101313

[106] Upadhyay, N., & Karmakar, A. (2012). The spectral subtractive-type algorithms for enhancing speech in noisy environments. *2012 1st International Conference on Recent*

Advances in Information Technology (RAIT), 841–847. https://doi.org/10.1109/RAIT.2012.6194534

[107] Valentini-Botinhao, Cassia. (2017). *Noisy speech database for training speech enhancement algorithms and TTS models* [Video recording]. School of Informatics. Centre for Speech Technology Research (CSTR). https://doi.org/https://doi.org/10.7488/ds/2117

[108] Venkateswarlu, S. C., Kumar, N. U., & Karthik, A. (2021). Speech Enhancement Using Recursive Least Square Based on Real-time adaptive filtering algorithm. *2021 6th International Conference for Convergence in Technology (I2CT)*, 1–4. https://doi.org/10.1109/I2CT51068.2021.9417929

[109] Vihari, S., Murthy, A. S., Soni, P., & Naik, D. C. (2016). Comparison of Speech Enhancement Algorithms. *Procedia Computer Science*, *89*, 666–676. https://doi.org/10.1016/j.procs.2016.06.032

[110] Wang, L., Zheng, W., Ma, X., & Lin, S. (2021). Denoising Speech Based on Deep Learning and Wavelet Decomposition. *Scientific Programming*, *2021*, 1–10. https://doi.org/10.1155/2021/8677043

[111] Wang, N. Y.-H., Wang, H.-L. S., Wang, T.-W., Fu, S.-W., Lu, X., Wang, H.-M., & Tsao, Y. (2021). Improving the Intelligibility of Speech for Simulated Electric and Acoustic Stimulation Using Fully Convolutional Neural Networks. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *29*, 184–195. https://doi.org/10.1109/TNSRE.2020.3042655

[112] Wang, P., Tan, K., & Wang, D. L. (2020). Bridging the Gap Between Monaural Speech Enhancement and Recognition With Distortion-Independent Acoustic Modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, 39–48. https://doi.org/10.1109/TASLP.2019.2946789

[113] Wijayasingha, L., & Stankovic, J. A. (2021). Robustness to noise for speech emotion classification using CNNs and attention mechanisms. *Smart Health*, *19*, 100165. https://doi.org/10.1016/j.smhl.2020.100165

[114] Williamson, D. S., & Wang, D. (2017). Time-Frequency Masking in the Complex Domain for Speech Dereverberation and Denoising. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(7), 1492–1501. https://doi.org/10.1109/TASLP.2017.2696307

[115] Williamson, D. S., Wang, Y., & Wang, D. (2016). Complex Ratio Masking for Monaural Speech Separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *24*(3), 483–492. https://doi.org/10.1109/TASLP.2015.2512042

[116] Wood, S. U. N., Stahl, J. K. W., & Mowlaee, P. (2019). Binaural Codebook-Based Speech Enhancement With Atomic Speech Presence Probability. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *27*(12), 2150–2161. https://doi.org/10.1109/TASLP.2019.2937174

[117] Xiang, Y., & Bao, C. (2020). A Parallel-Data-Free Speech Enhancement Method Using Multi-Objective Learning Cycle-Consistent Generative Adversarial Network. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, 1826–1838. https://doi.org/10.1109/TASLP.2020.2997118

[118] Xiao, X., & Nickel, R. M. (2010). Speech Enhancement With Inventory Style Speech Resynthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*(6), 1243–1257. https://doi.org/10.1109/TASL.2009.2031793

[119] Xie, S., & Li, L. (2021). Improvement and Application of Deep Belief Network Based on Sparrow Search Algorithm. *2021 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA)*, 705–708. https://doi.org/10.1109/AEECA52519.2021.9574138

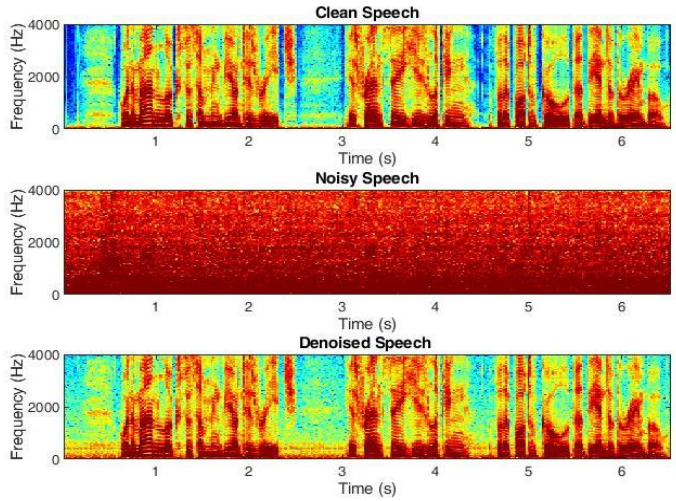
[120] Xu, Y., Du, J., Dai, L.-R., & Lee, C.-H. (2015a). A Regression Approach to Speech Enhancement Based on Deep Neural Networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *23*(1), 7–19. https://doi.org/10.1109/TASLP.2014.2364452

[121] Yang, H., Choe, S., Kim, K., & Kang, H.-G. (2018). Deep learning-based speech presence probability estimation for noise PSD estimation in single-channel speech enhancement. *2018 International Conference on Signals and Systems (ICSigSys)*, 267–270. https://doi.org/10.1109/ICSIGSYS.2018.8372770

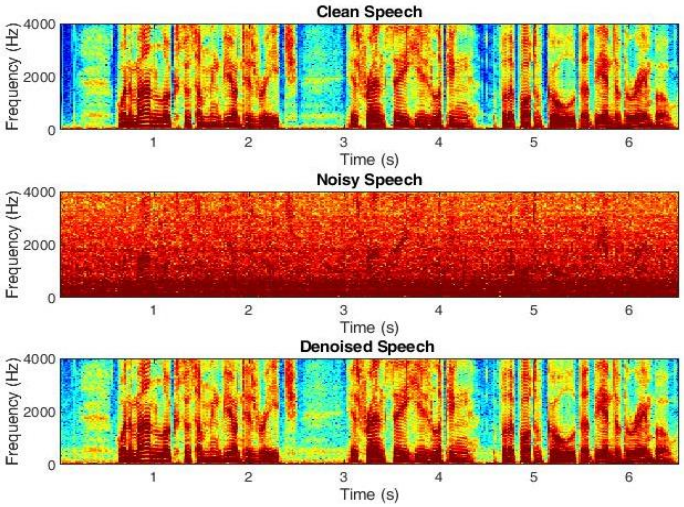
[122] Yuan, W. (2020). A time–frequency smoothing neural network for speech enhancement. *Speech Communication*, *124*, 75–84. https://doi.org/10.1016/j.specom.2020.09.002

[123] Zhang, Q., & Wang, M. (2017). Speech enhancement for nonstationary noise environments. *2017 IEEE 17th International Conference on Communication Technology (ICCT)*, 1663–1667. https://doi.org/10.1109/ICCT.2017.8359913

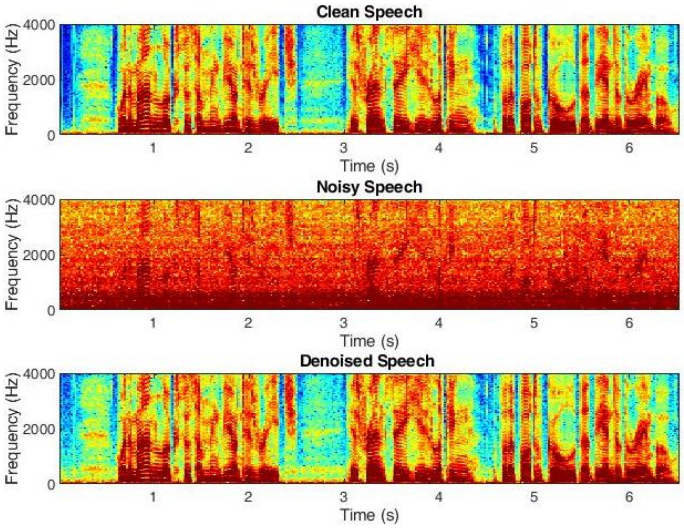
APPENDIX 1 (Spectrograms of DFNN)



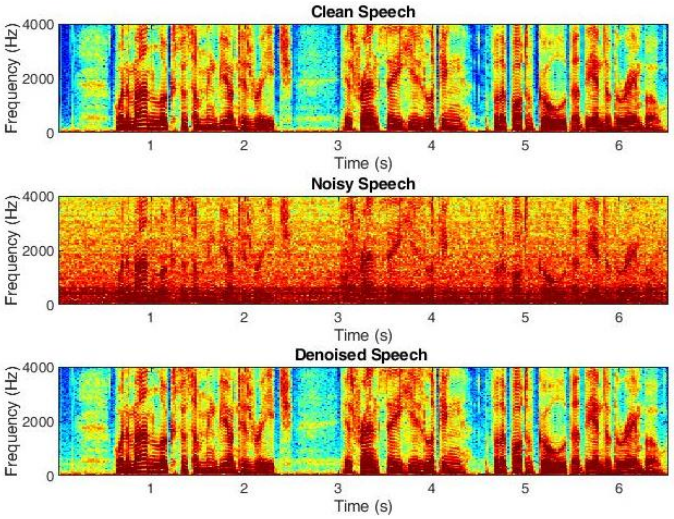
a) -10 dB



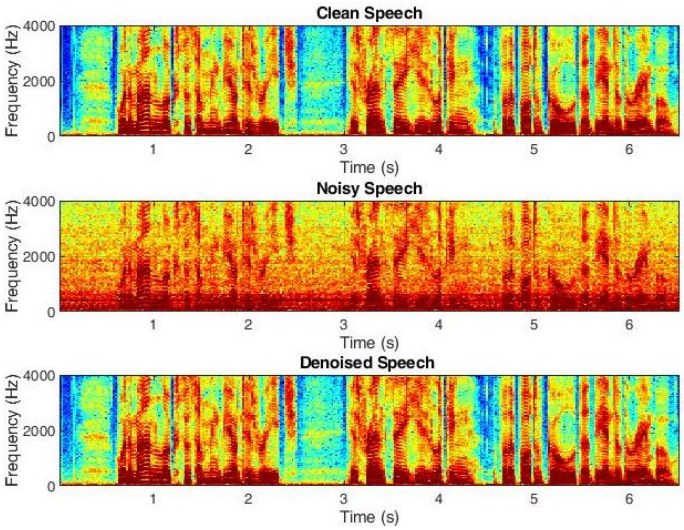
b) -5 dB



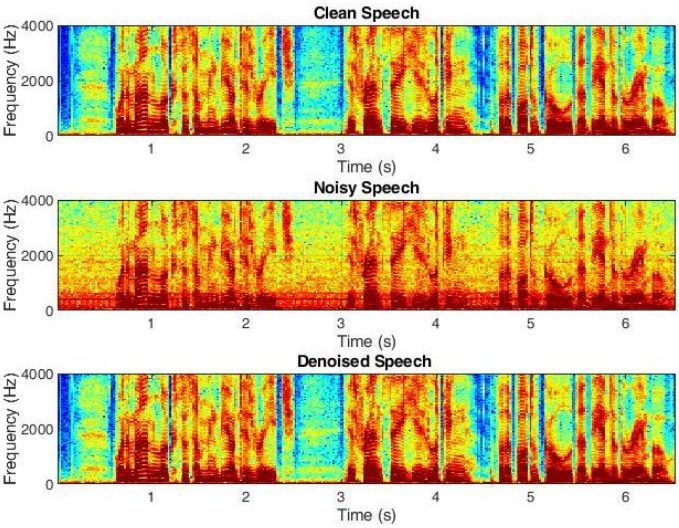
c) 0 dB



d) 5 dB

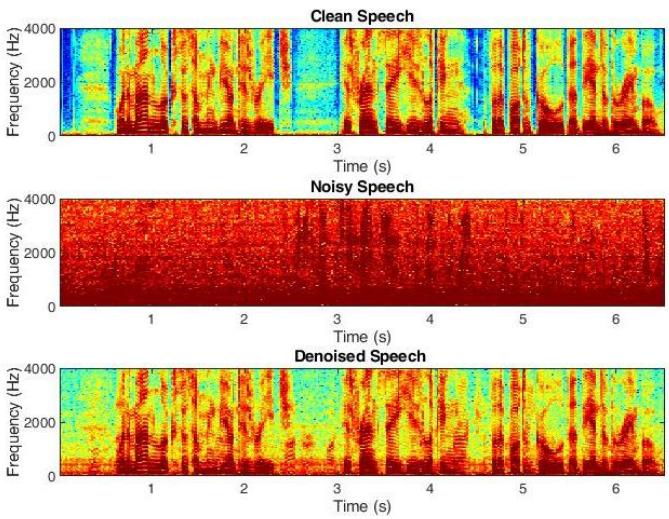


e) 10 dB

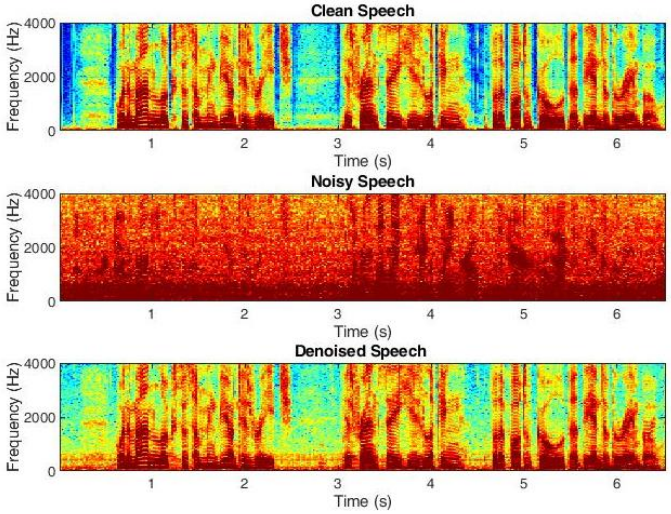


f) 15 dB

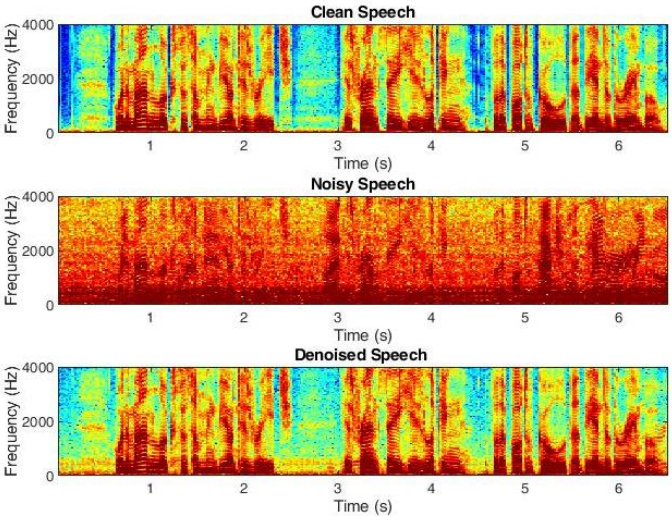
Figure 1 DFNN - Spectrogram Images of Washing Machine Noise for various Noise Levels



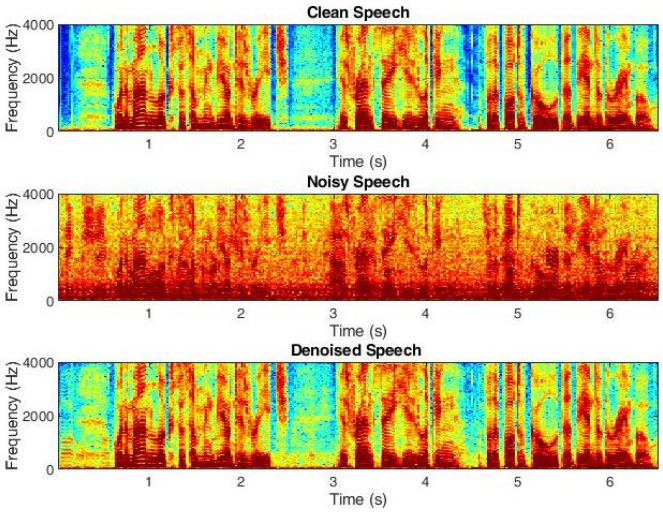
a) -10 dB



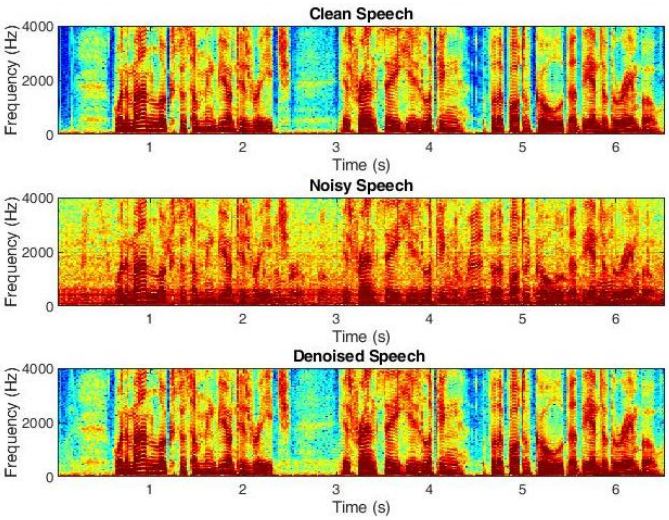
b) -5 dB



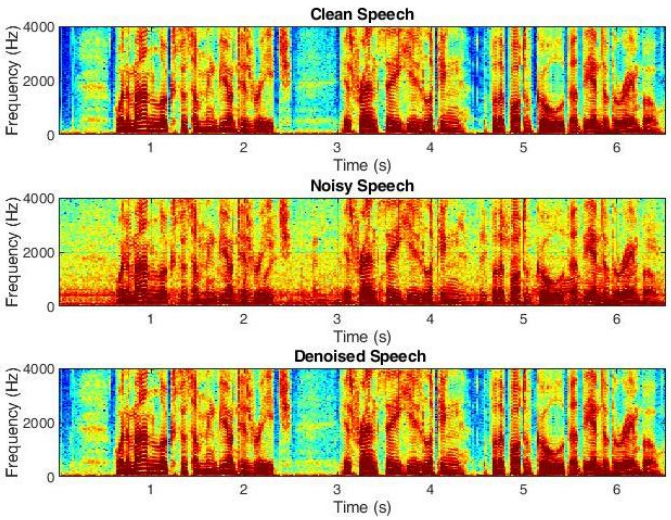
c) 0 dB



d) 5 dB

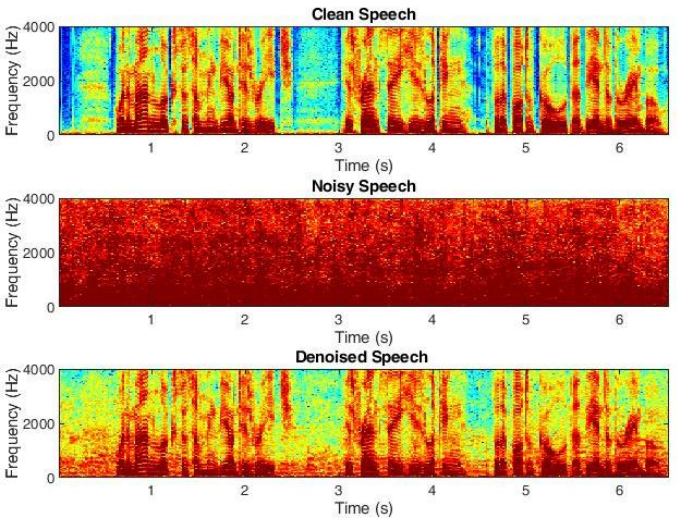


e) 10 dB

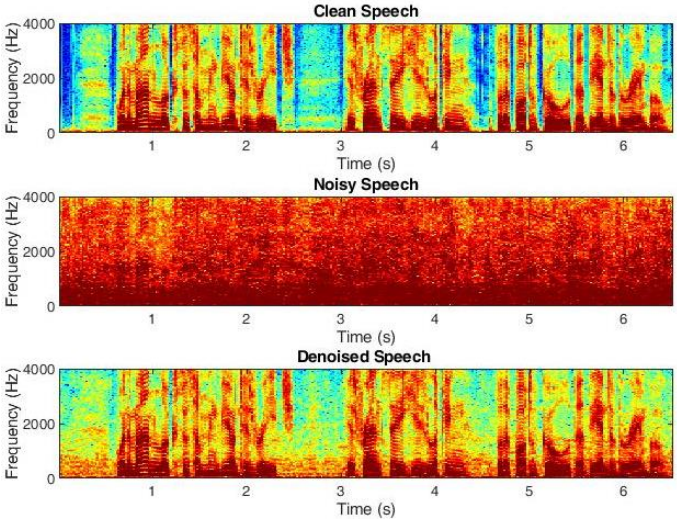


f) 15 Db

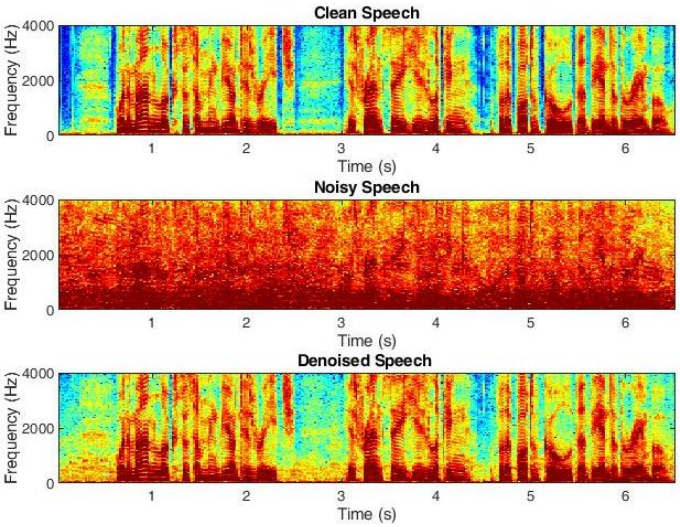
Figure 2 DFNN – Spectrogram Images of Rainbow Noise for various Noise Levels



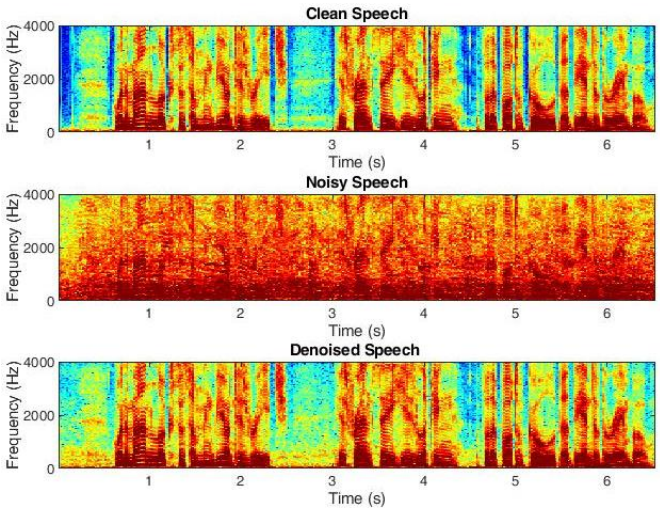
a) -10 dB



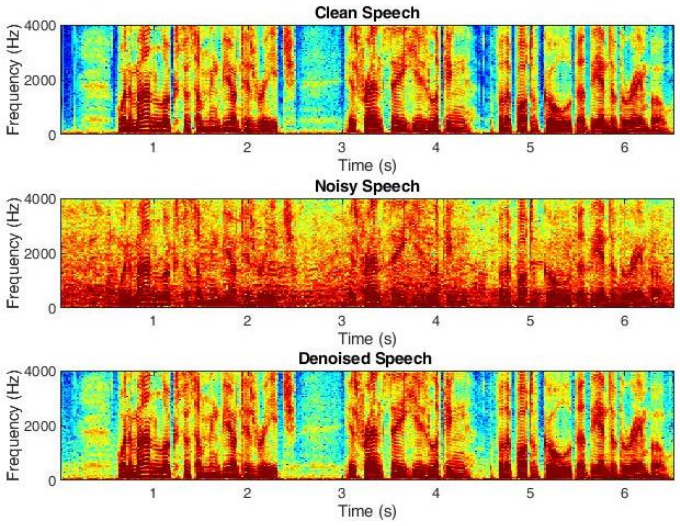
b) -5 dB



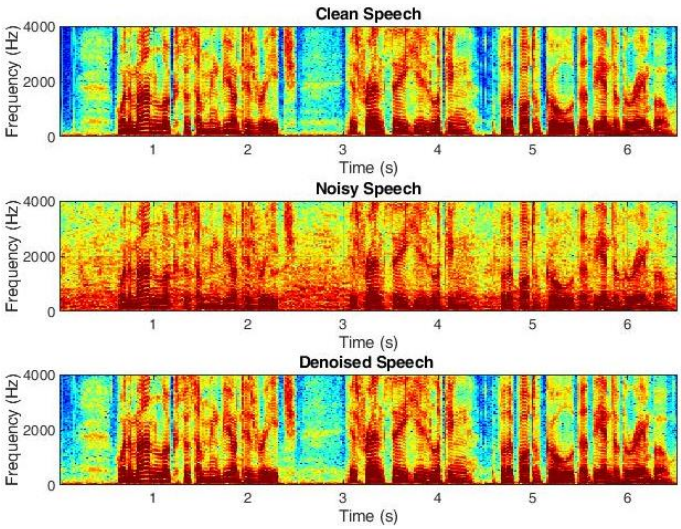
c) 0 dB



d) 5 dB

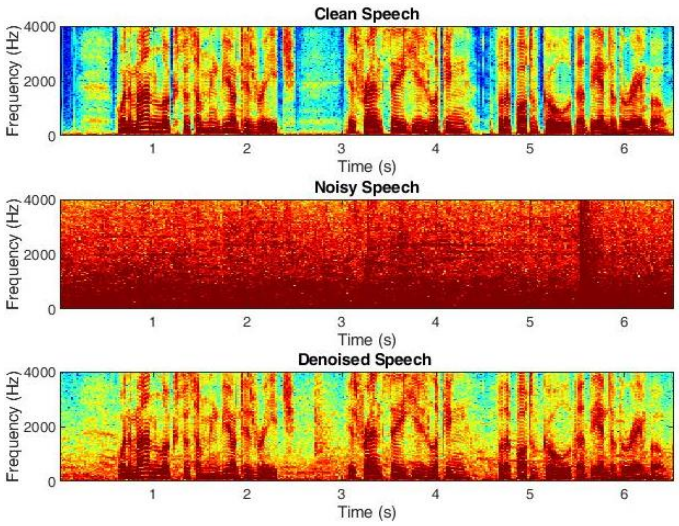


e) 10 dB

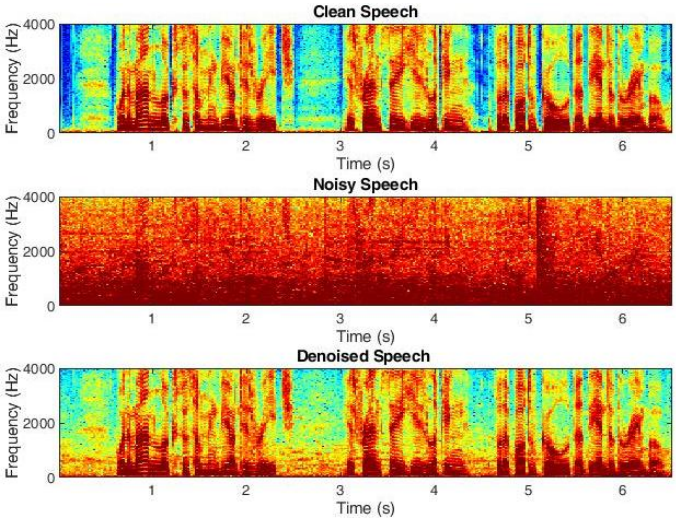


f) 15 dB

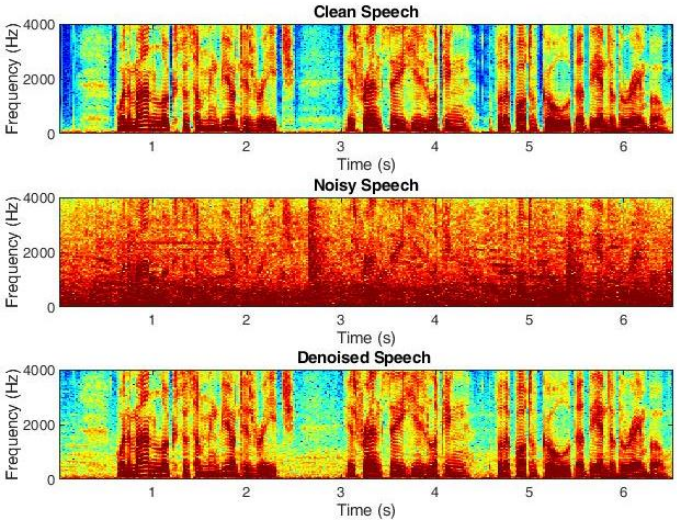
Figure 3 DFNN – Spectrogram Images of Babble Noise for various Noise Levels



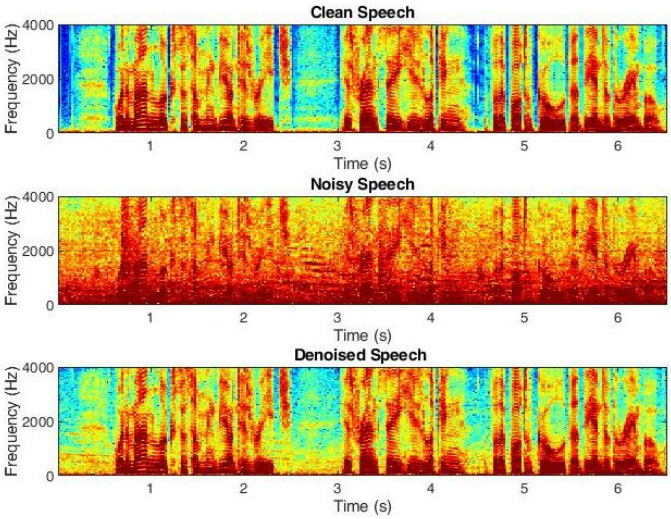
a) -10 dB



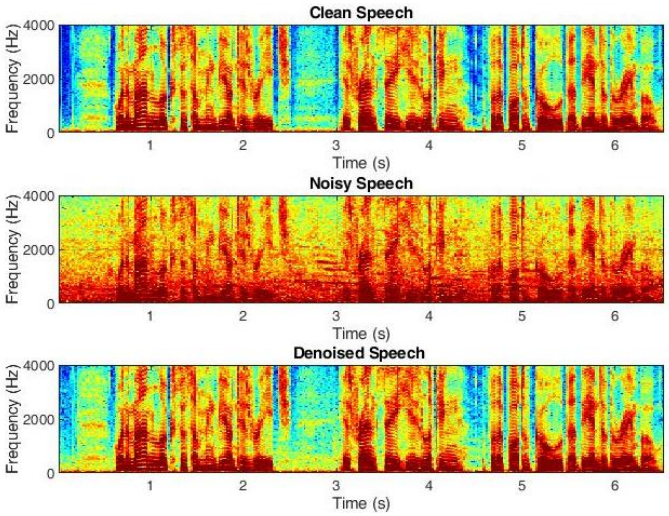
b) -5 dB



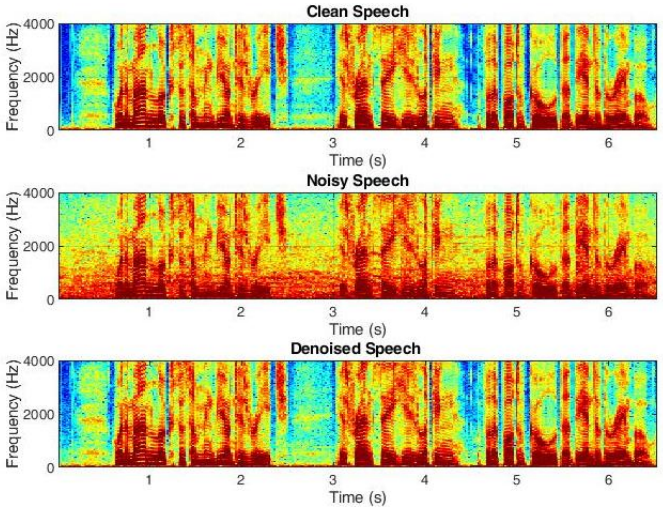
c) 0 dB



d) 5 dB



e) 10 dB



f) 15 dB

Figure 4 DFNN Spectrogram Images of Airport Noise for various Noise Levels

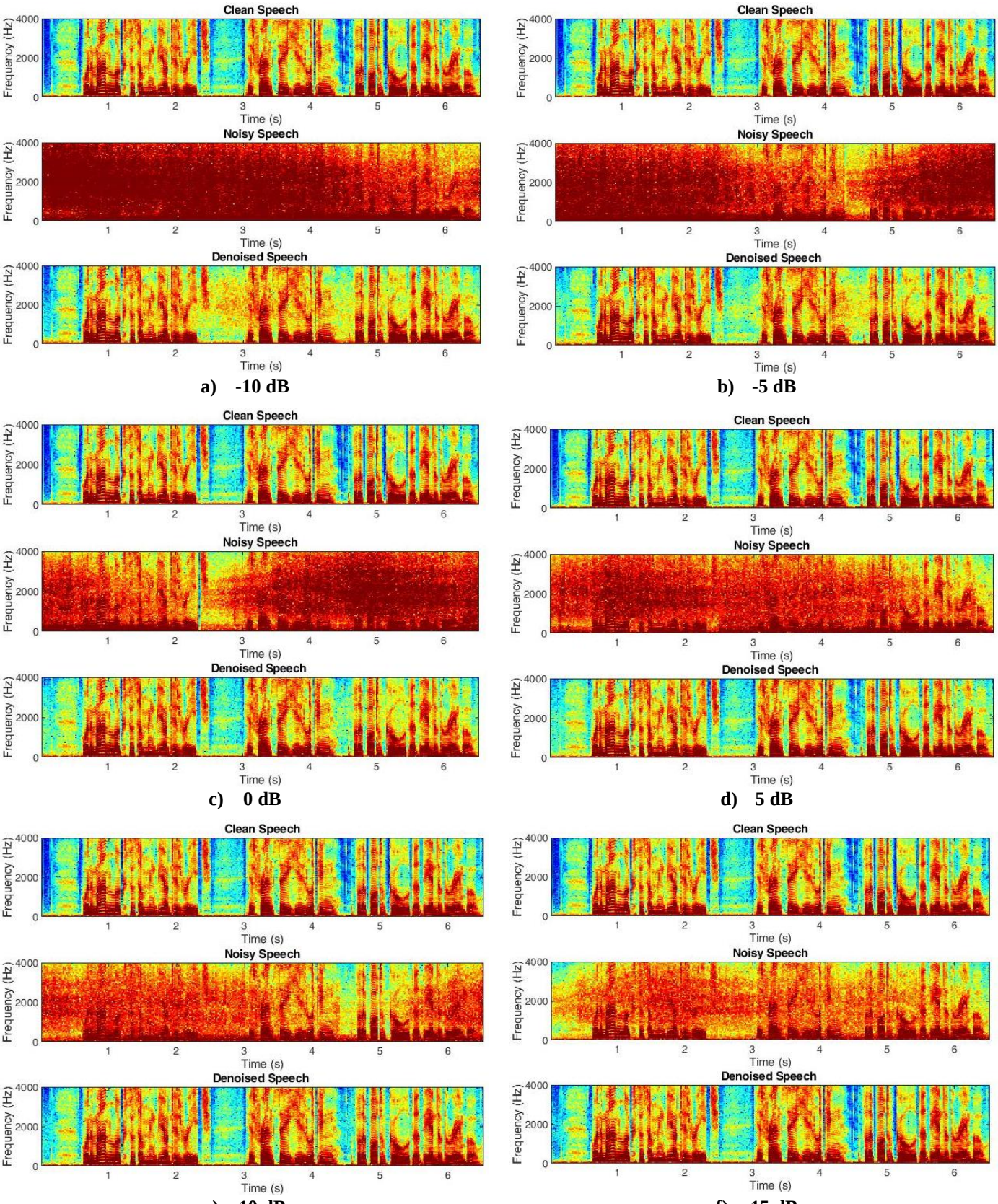


Figure 5 DFNN – Spectrogram Images of Jetplane Noise for various Noise Levels

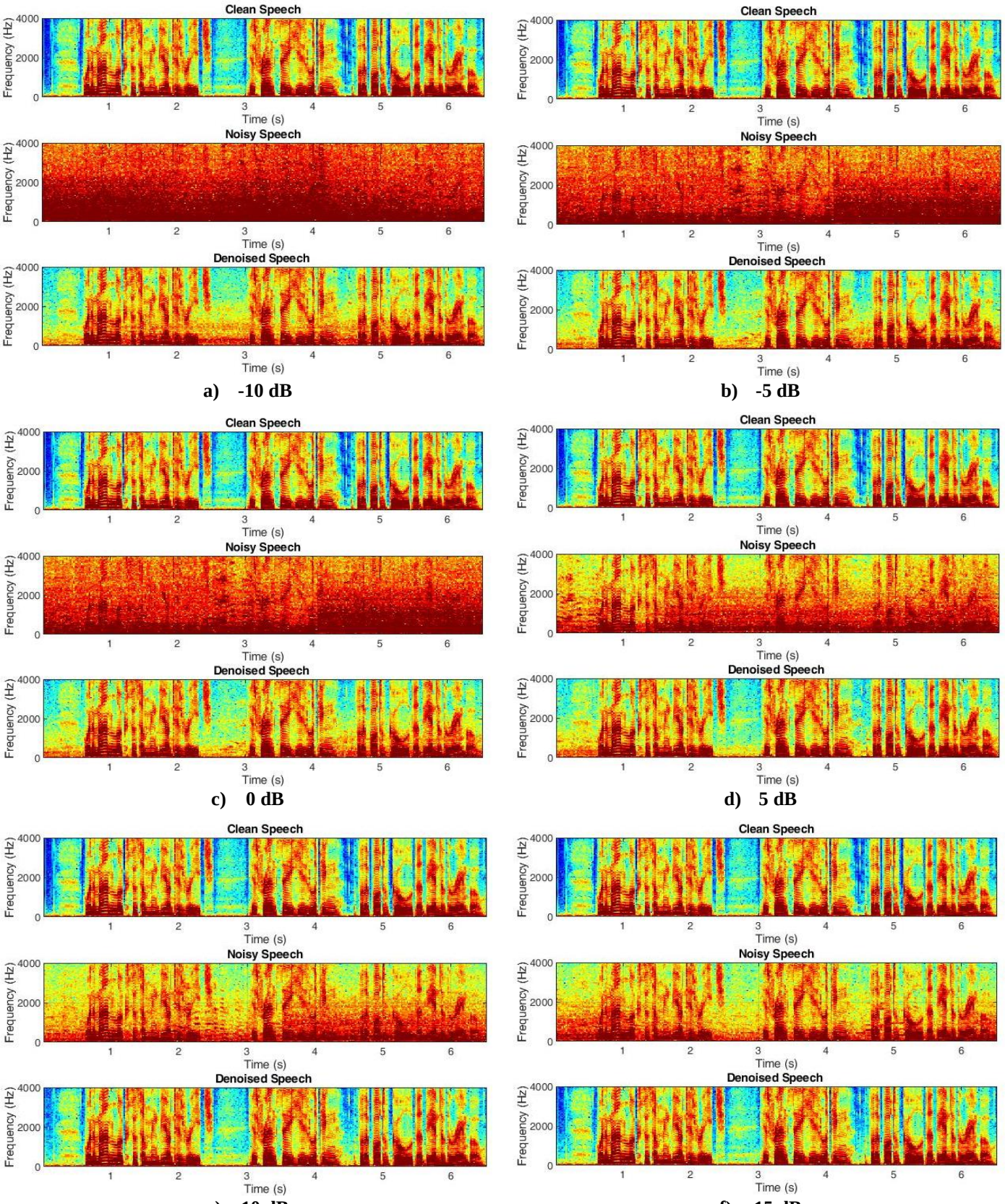
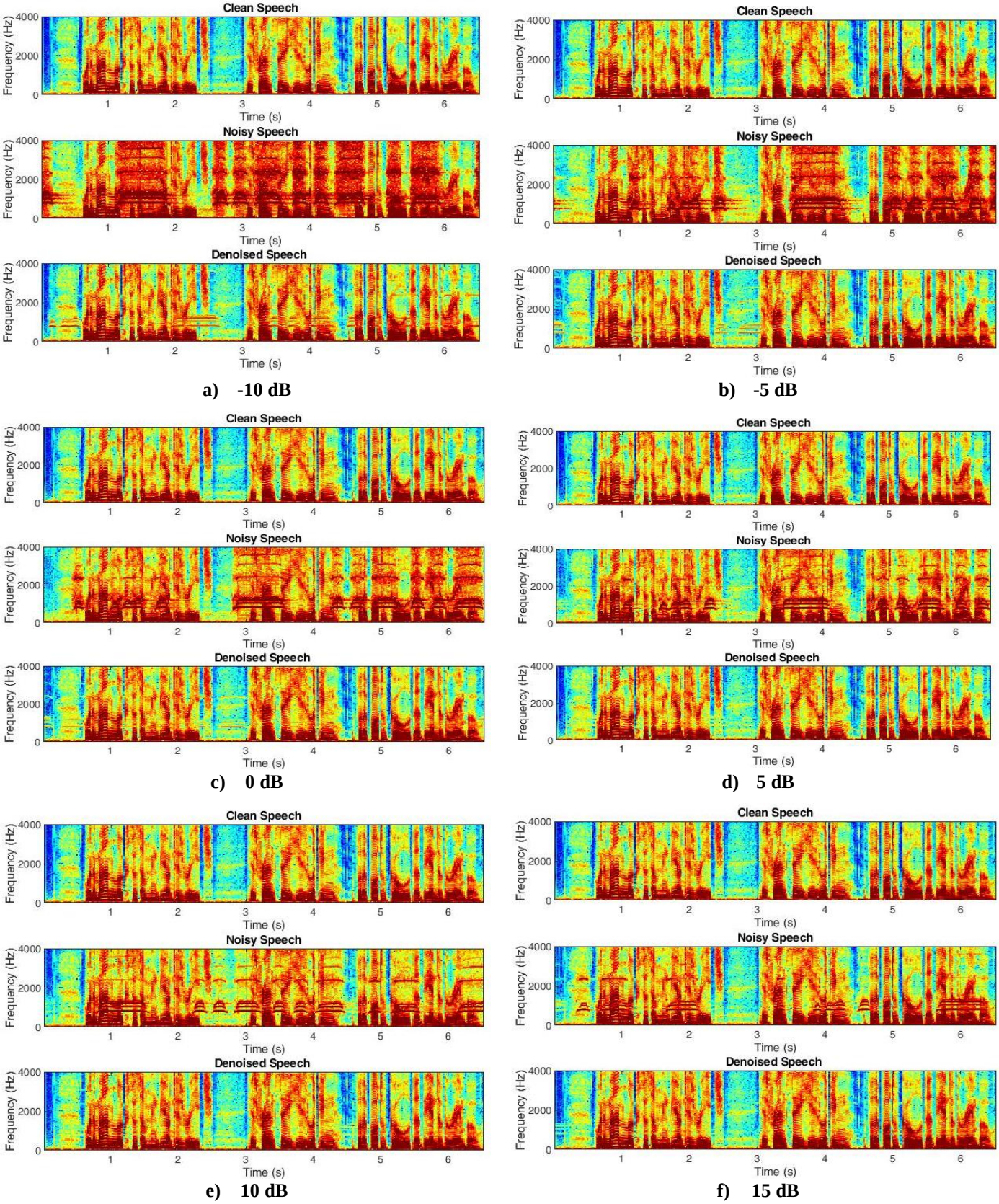


Figure 6 DFNN – Spectrogram Images of Street Noise for various Noise Levels

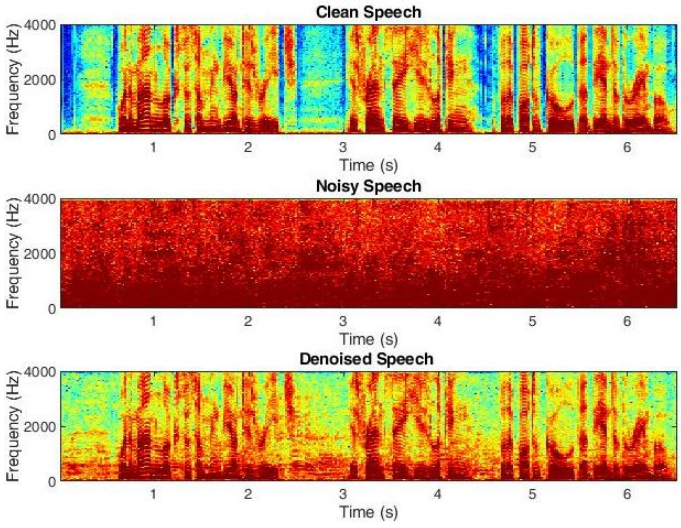


a) -10 dB b) -5 dB

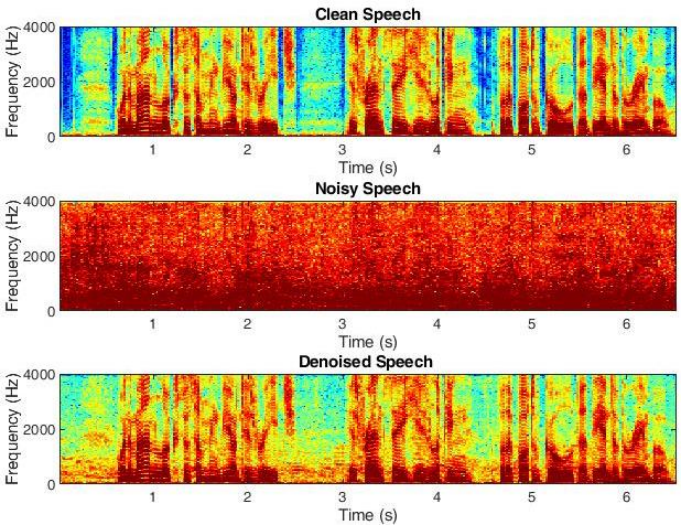
c) 0 dB d) 5 dB

e) 10 dB f) 15 dB

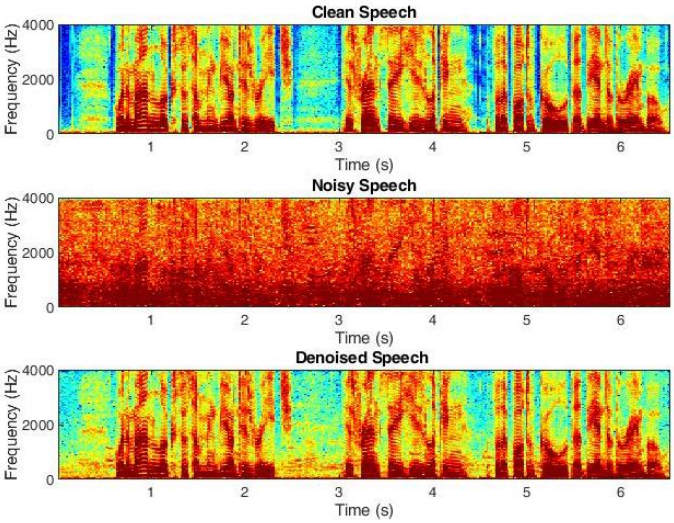
Figure 7 DFNN – Spectrogram Images of Train Whistle Noise for various Noise Levels



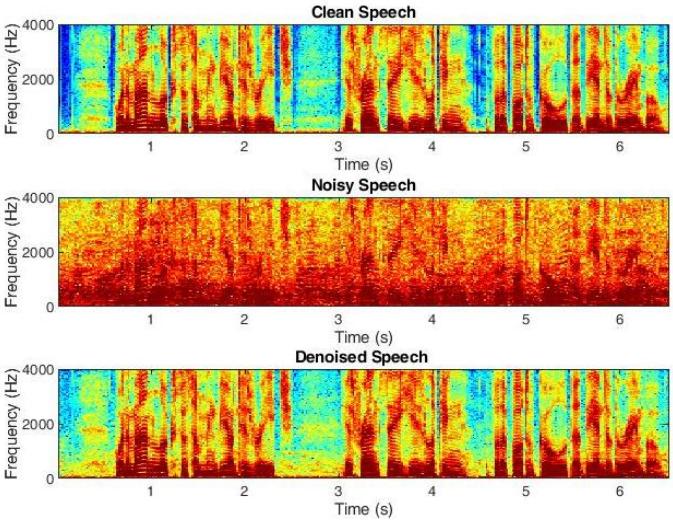
a) -10 dB



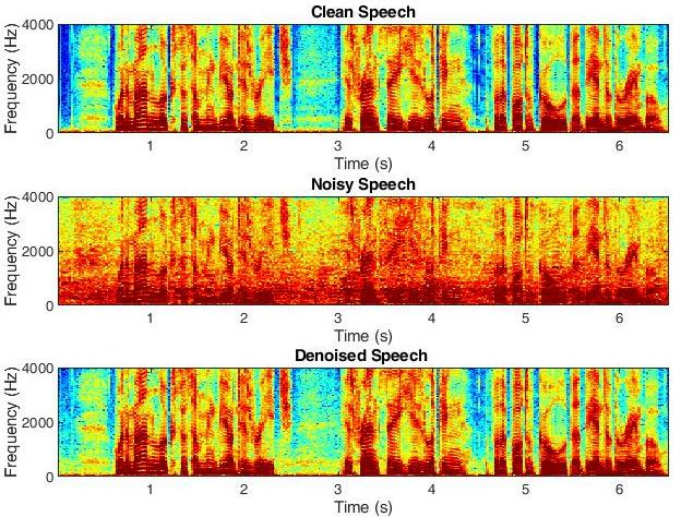
b) -5 dB



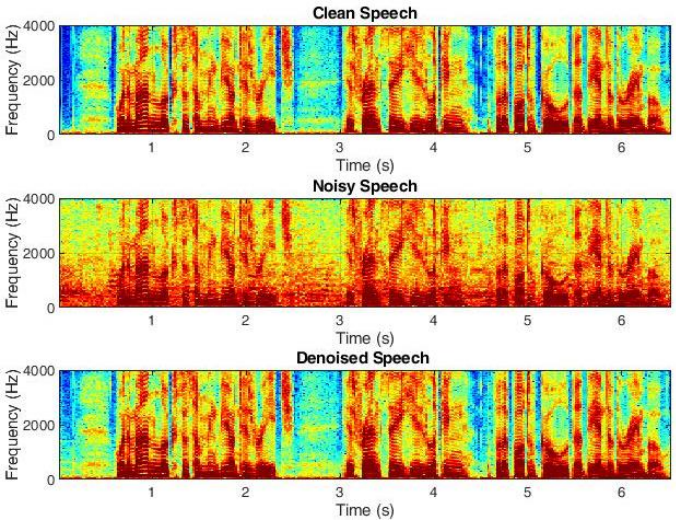
c) 0 dB



d) 5 dB

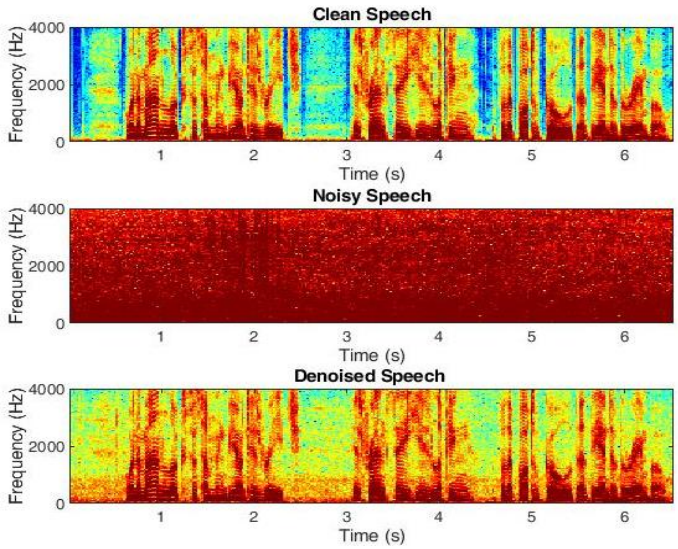


e) 10 dB

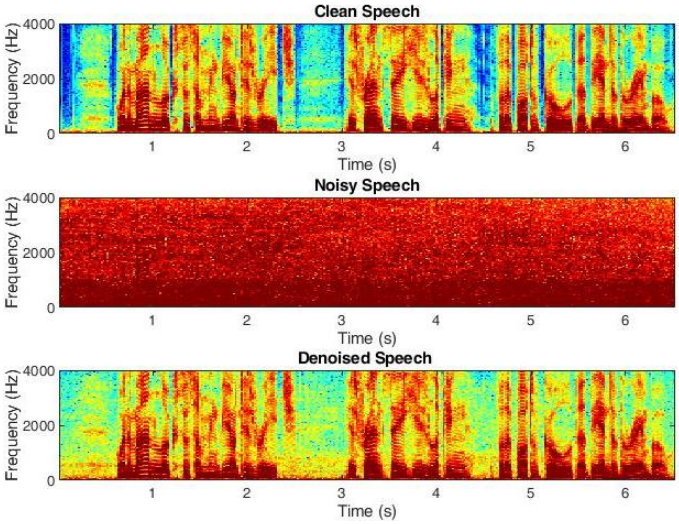


f) 15 dB

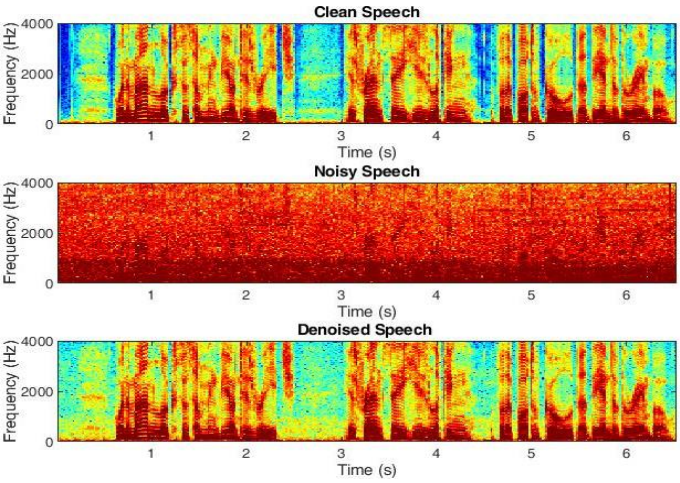
Figure 8 DFNN – Spectrogram Images of Restaurant Noise for various Noise Levels



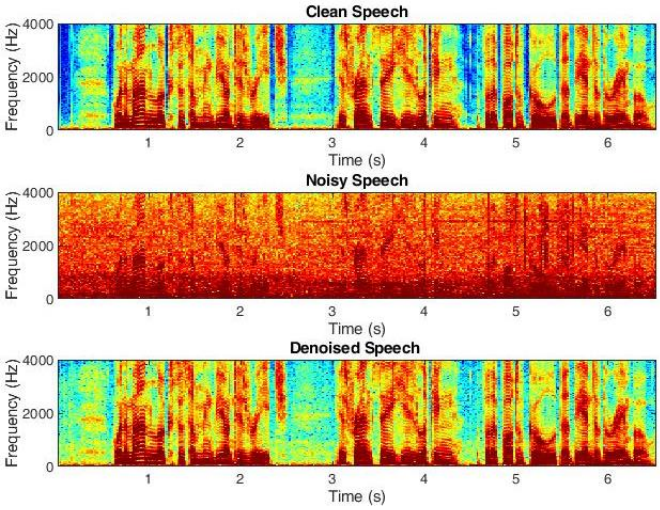
a) -10 dB



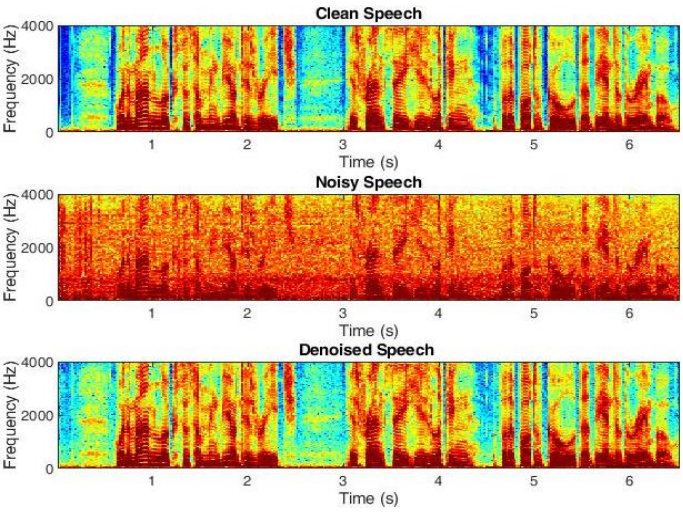
b) -5 dB



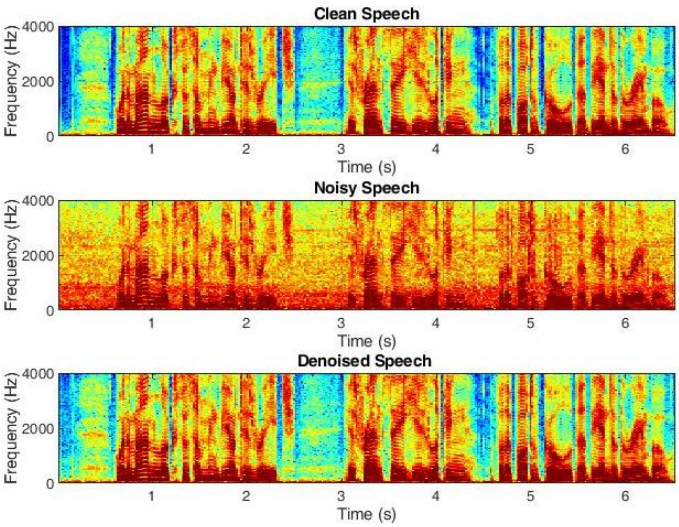
c) 0 dB



d) 5 dB



e) 10 dB



f) 15 dB

Figure 9 DFNN – Spectrogram Images of Car Noise for various Noise Levels

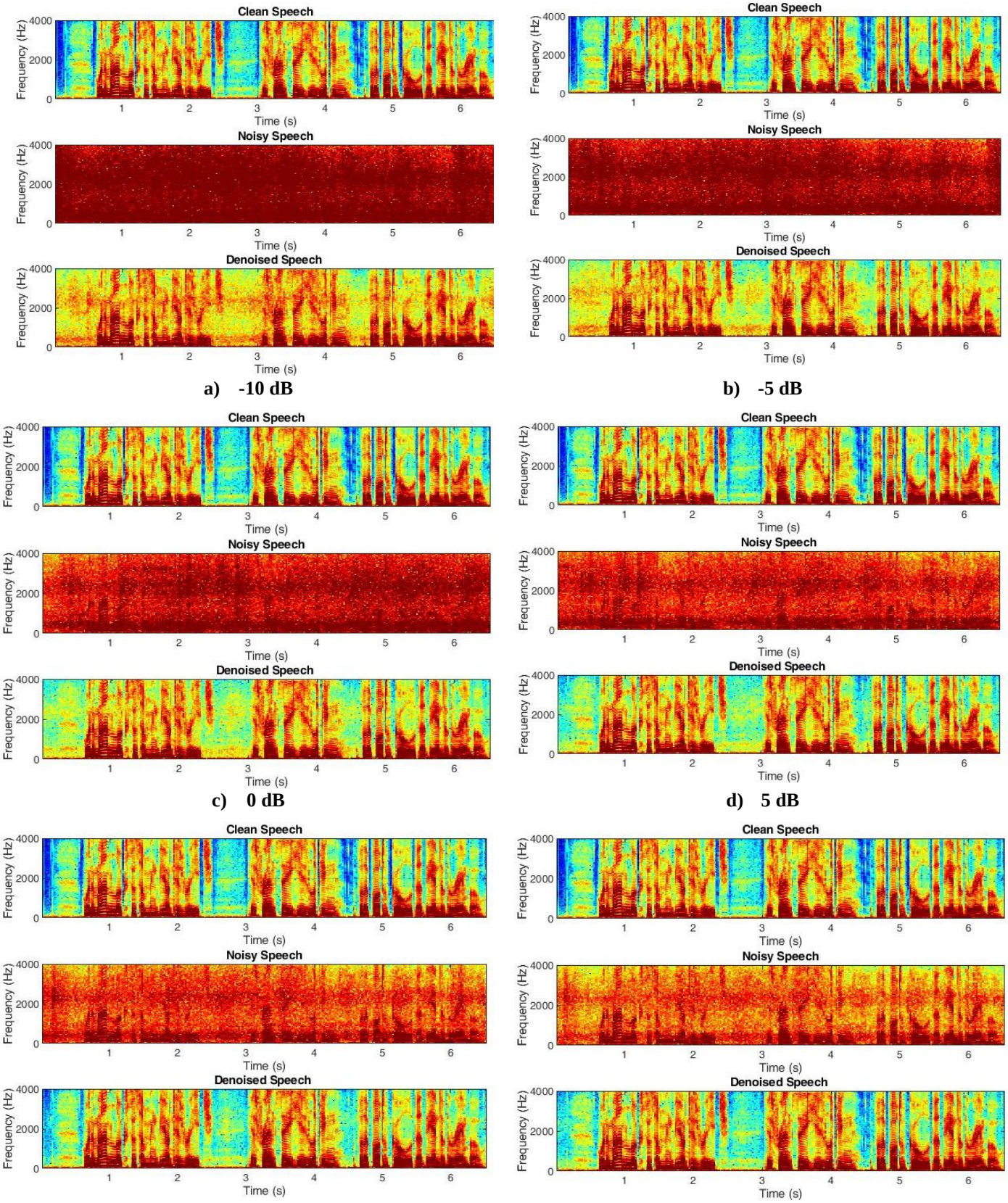
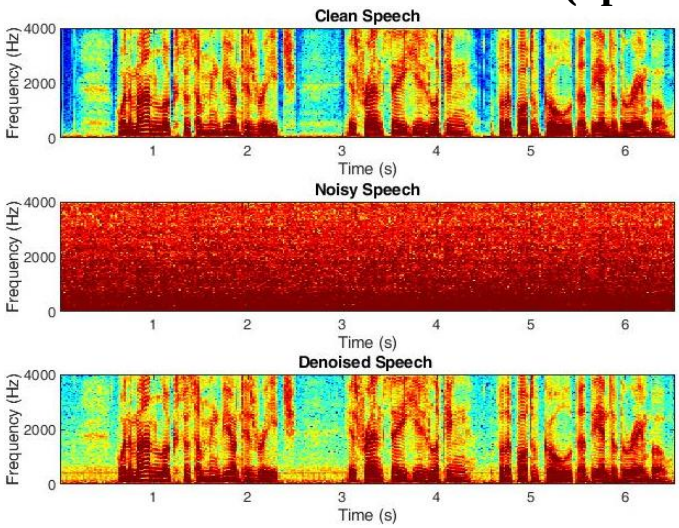
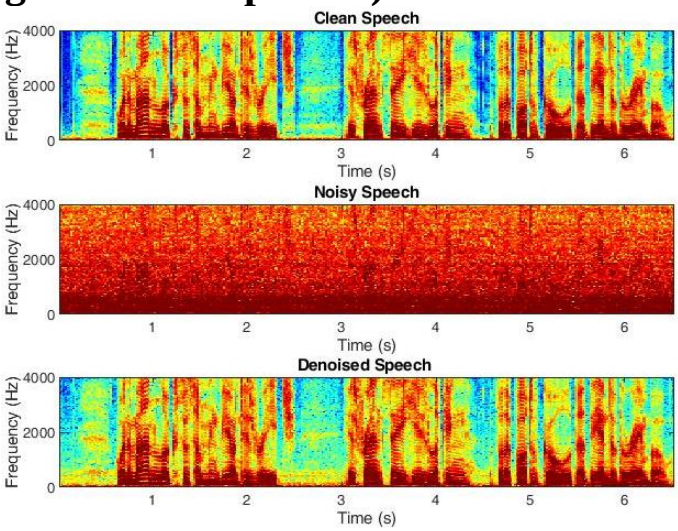


Figure 10 DFNN – Spectrogram Images of Subway Noise for various Noise Levels

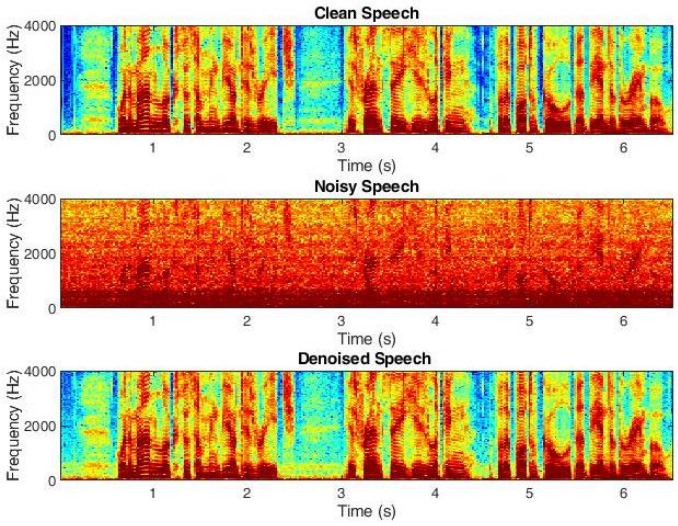
APPENDIX 2 (Spectrograms of Deep CNN)



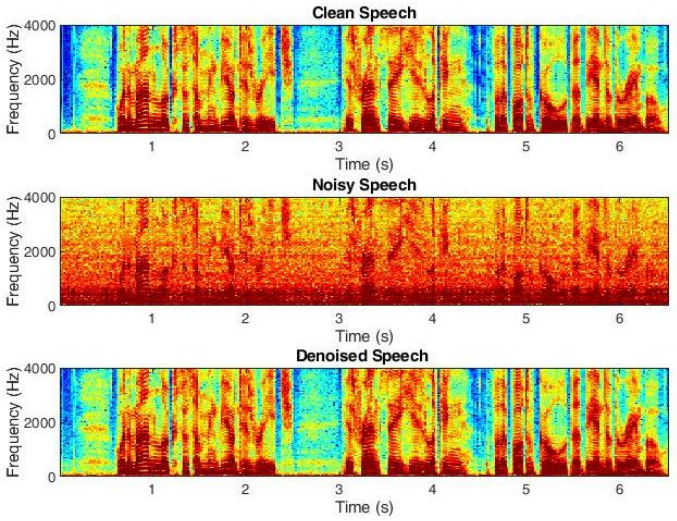
a) -10 dB



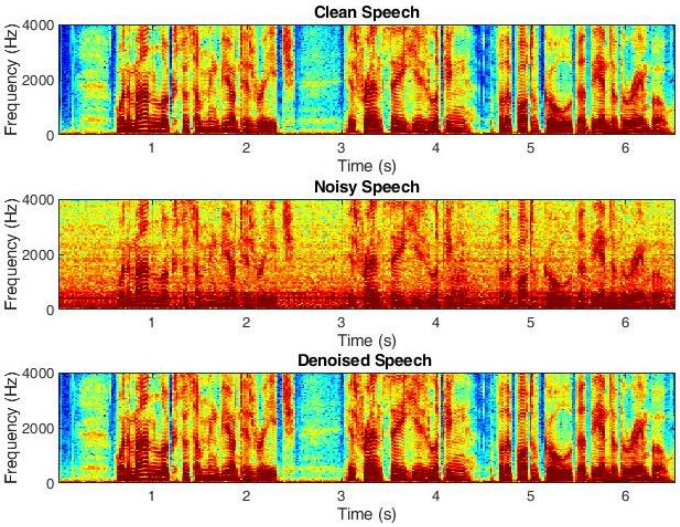
b) -5 dB



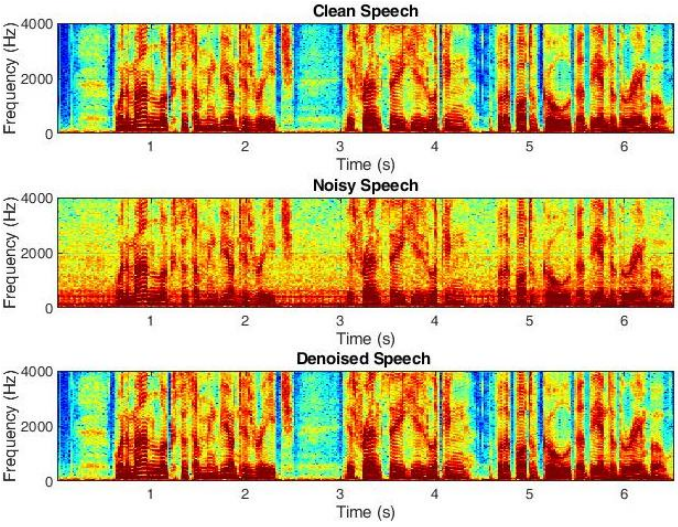
c) 0 dB



d) 5 dB

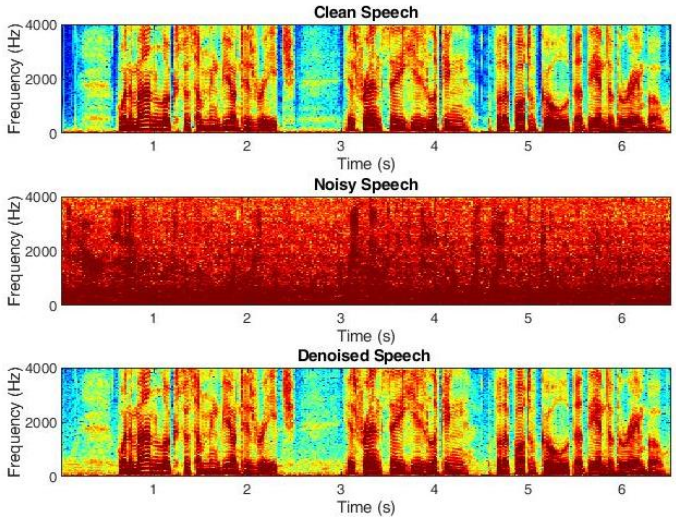


e) 10 dB

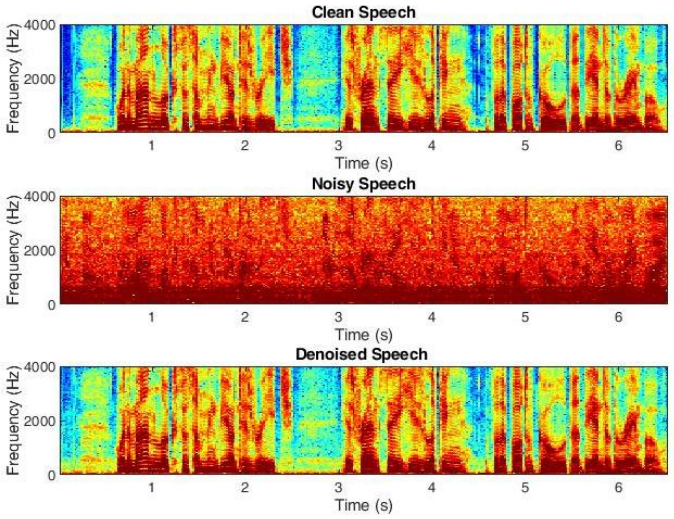


f) 15 dB

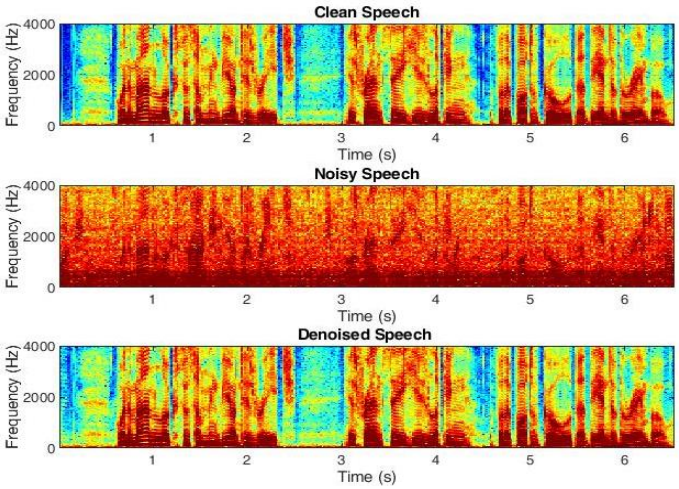
Figure 1 Deep CNN - Spectrogram Images of Washing Machine Noise for various Noise Levels



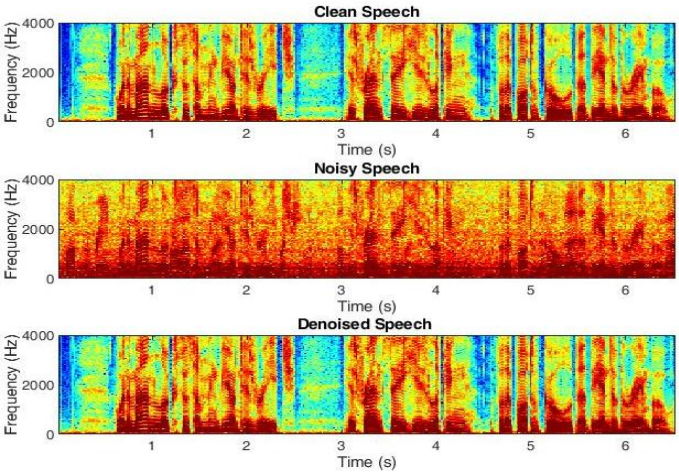
a) -10 dB



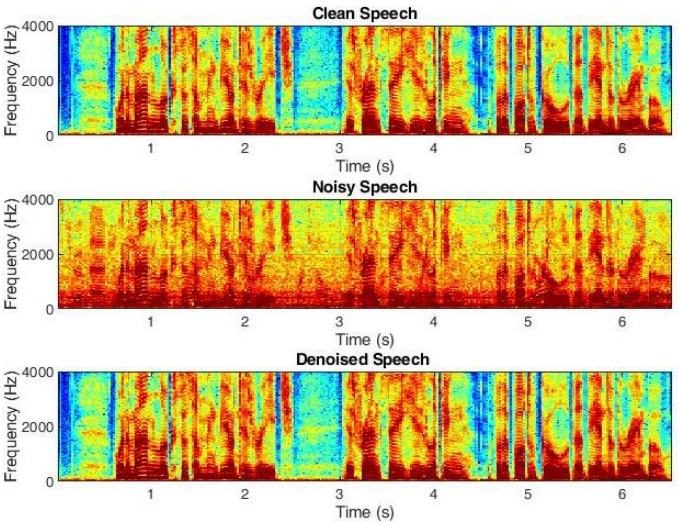
b) -5 dB



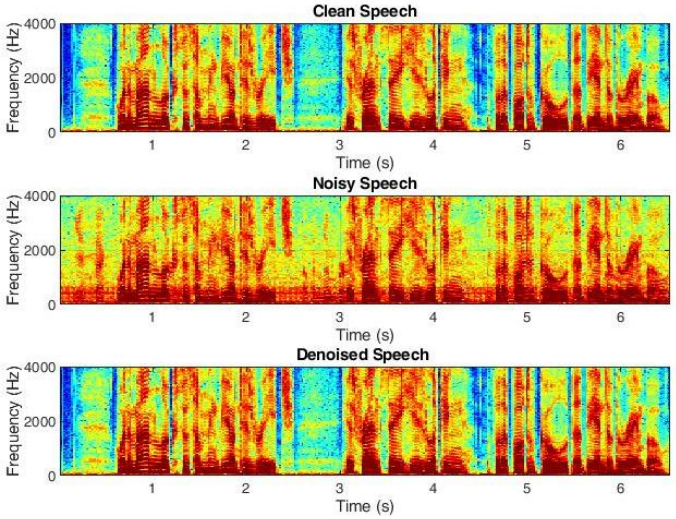
c) 0 dB



d) 5 dB

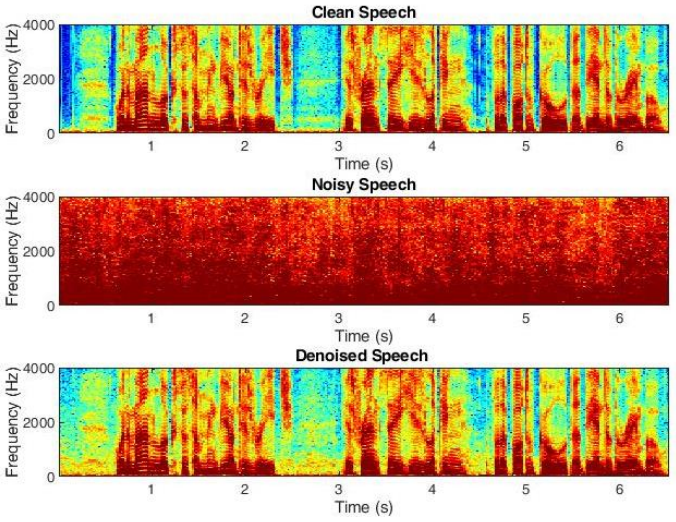


e) 10 dB

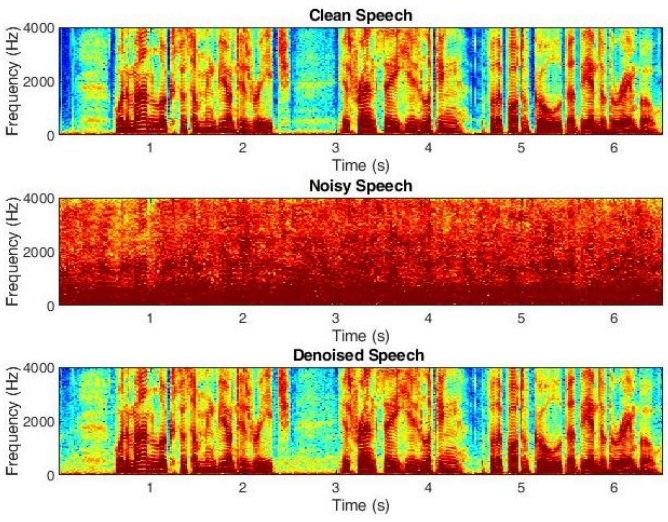


f) 15 dB

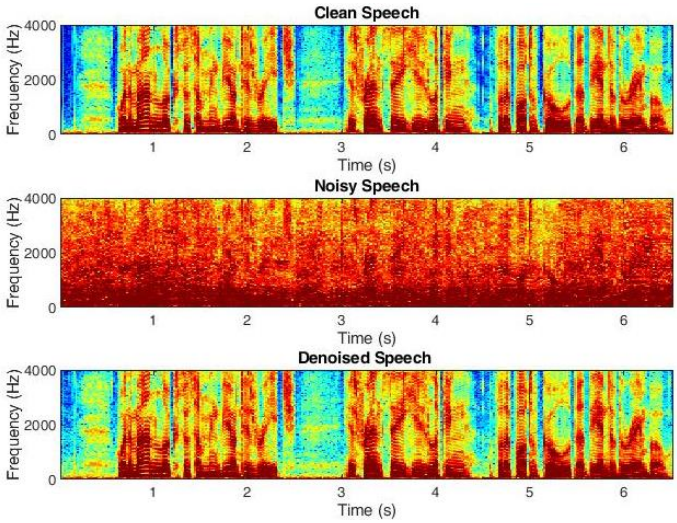
Figure 2 Deep CNN – Spectrogram Images of Rainbow Noise for various Noise Levels



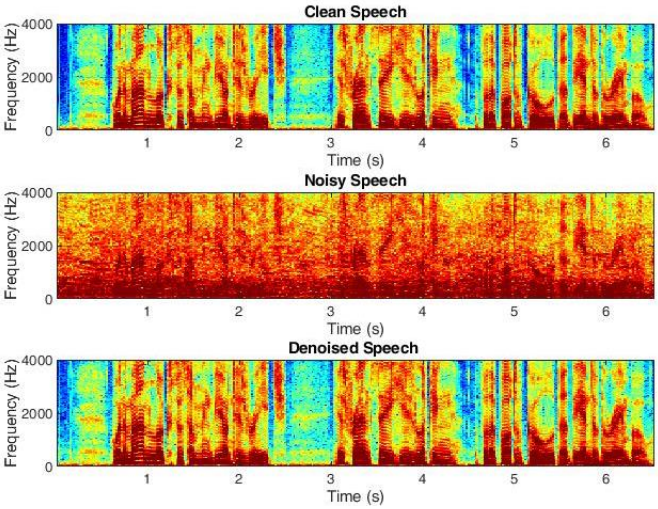
a) -10 dB



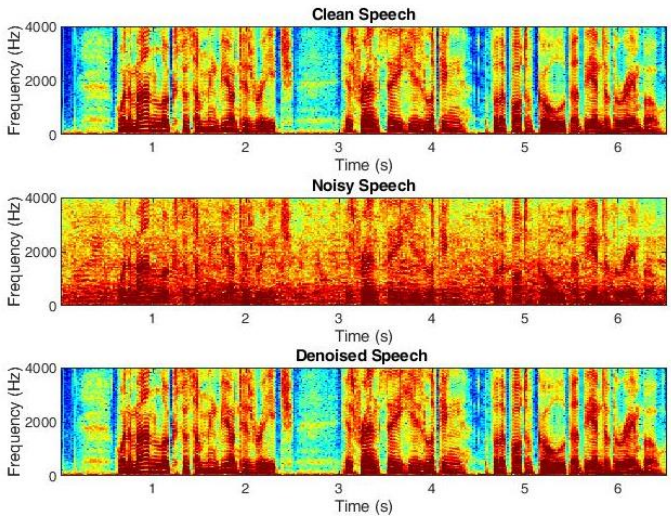
b) -5 dB



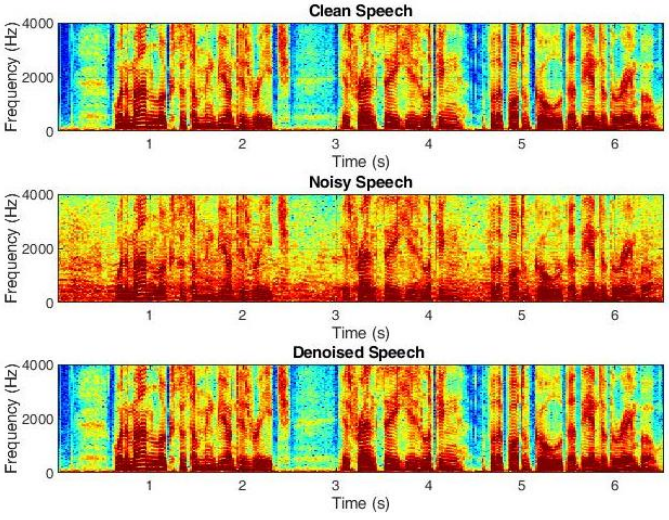
c) 0 dB



d) 5 dB

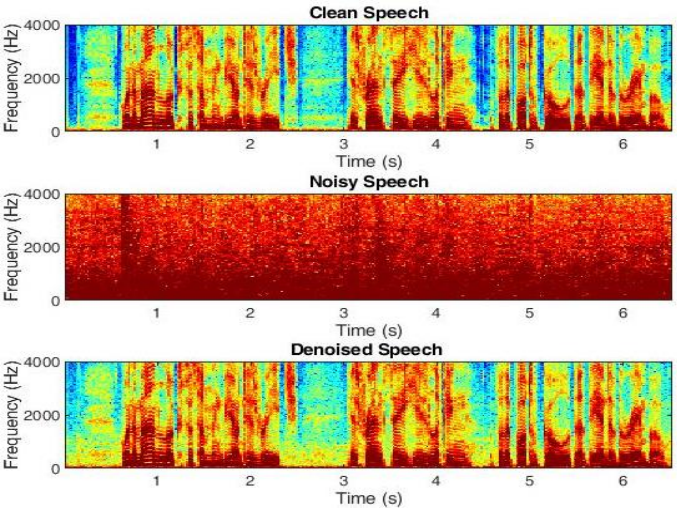


e) 10 dB

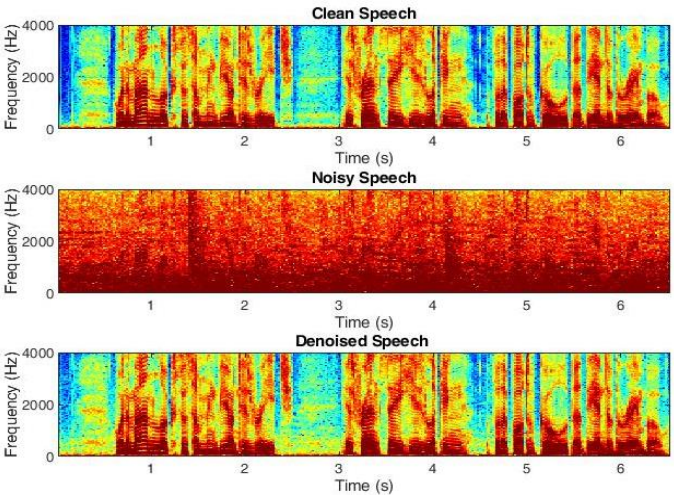


f) 15 dB

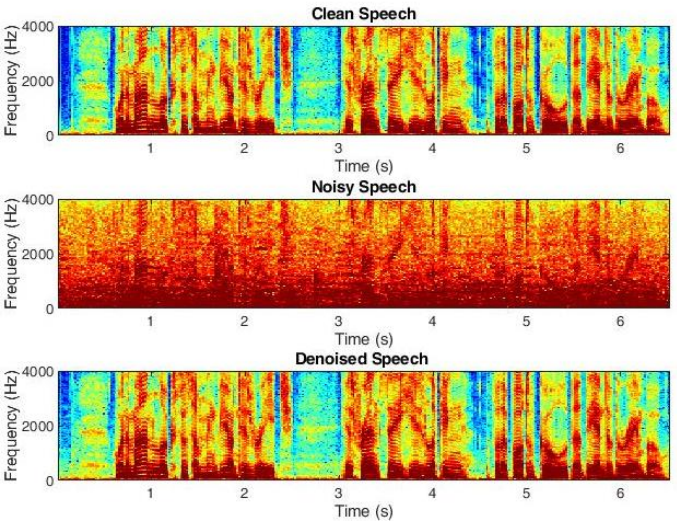
Figure 3 Deep CNN – Spectrogram Images of Babble Noise for various Noise Levels



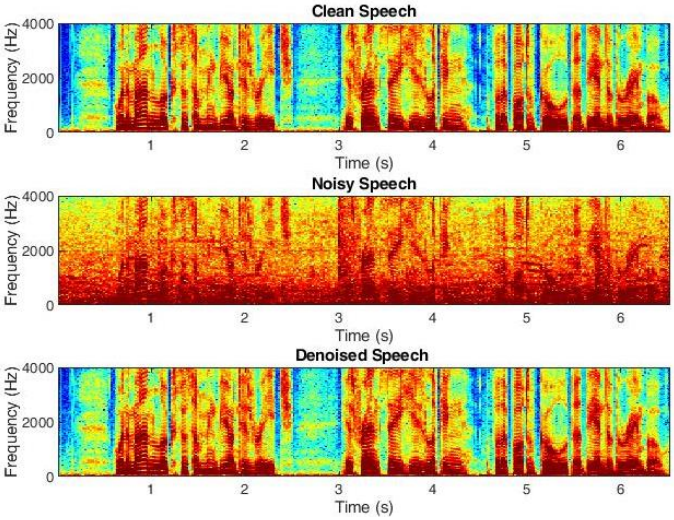
a) -10 dB



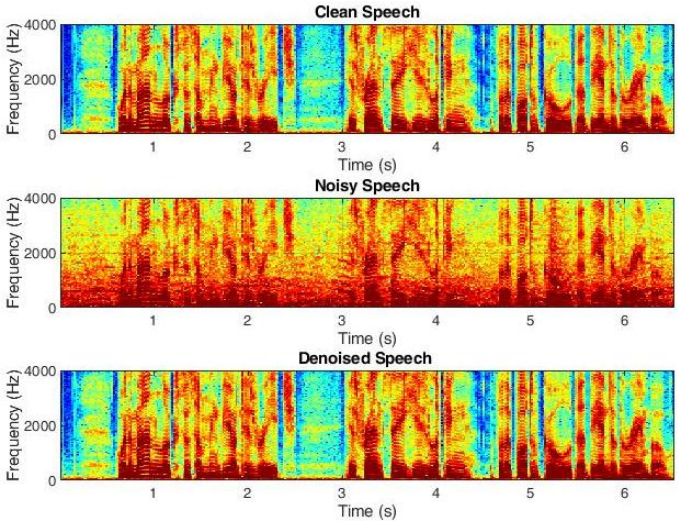
b) -5 dB



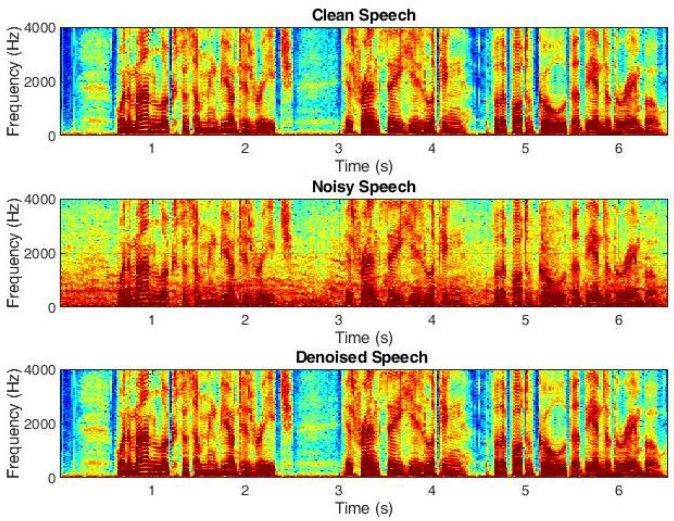
c) 0 dB



d) 5 dB



e) 10 dB



f) 15 dB

Figure 4 Deep CNN – Spectrogram Images of Airport Noise for various Noise Levels

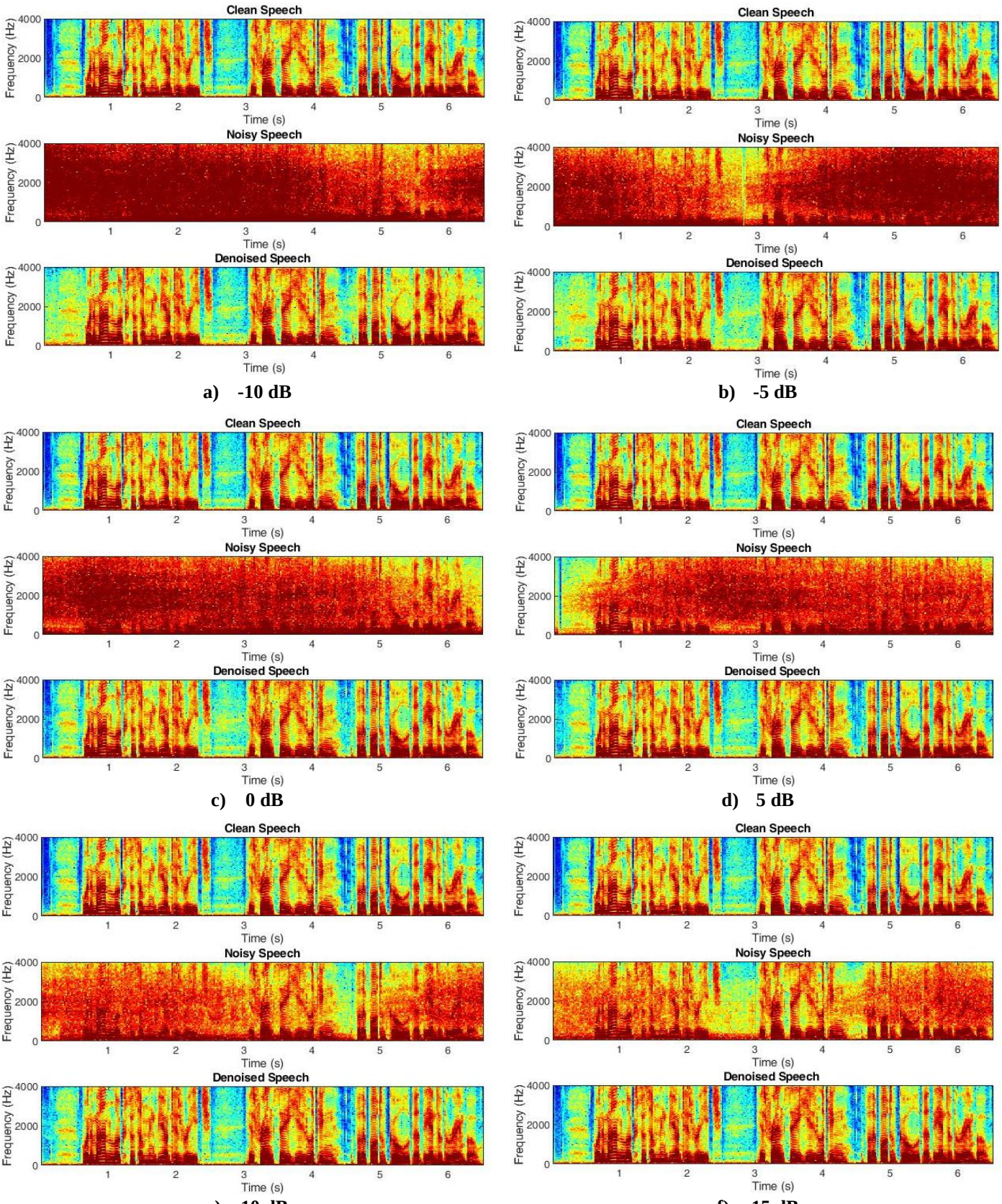
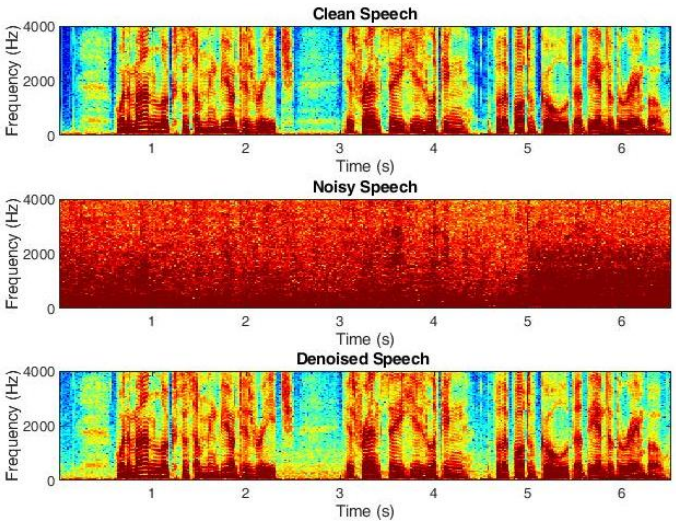
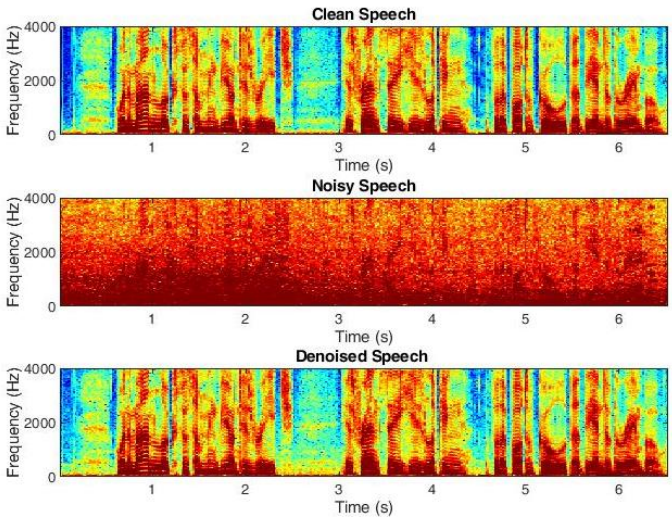


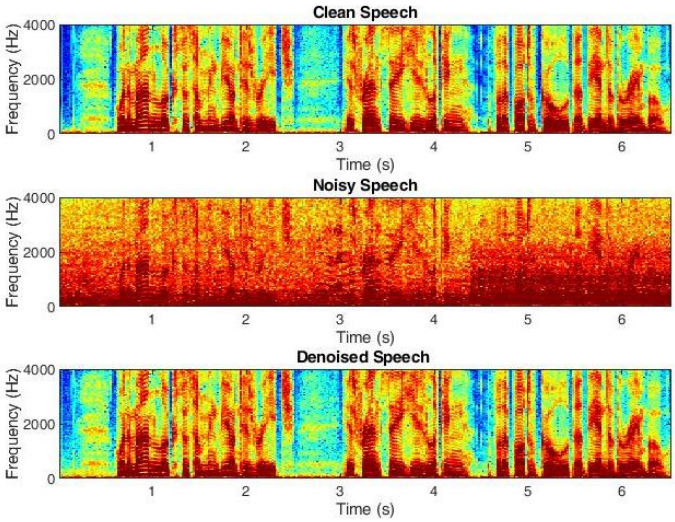
Figure 5 Deep CNN – Spectrogram Images of Jetplane Noise for various Noise Levels



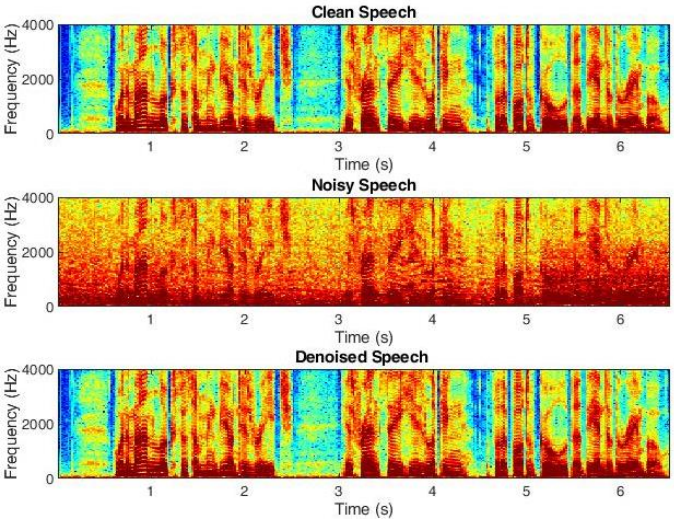
a) -10 dB



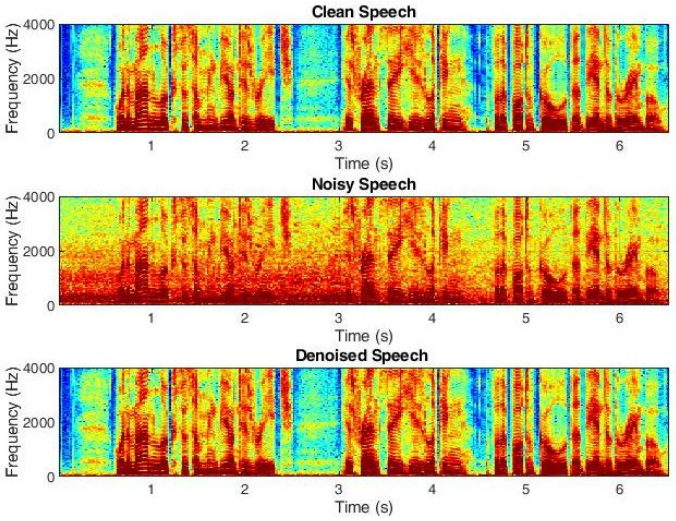
b) -5 dB



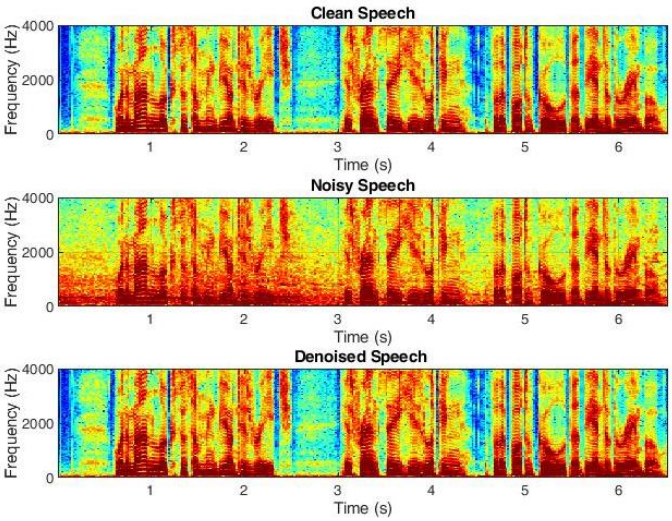
c) 0 dB



d) 5 dB



e) 10 dB



f) 15 dB

Figure 6 Deep CNN – Spectrogram Images of Street Noise for various Noise Levels

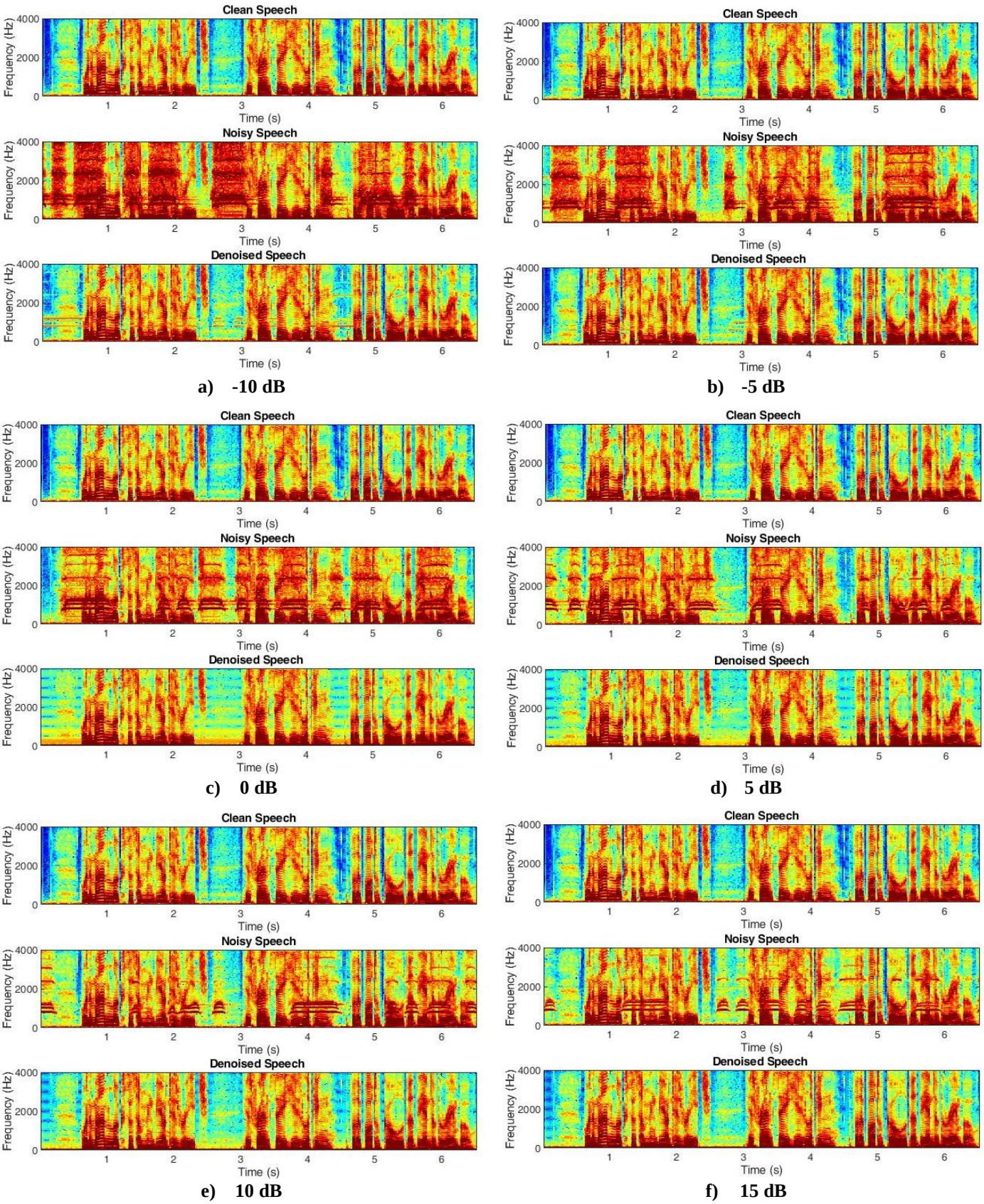


Figure 7 Deep CNN – Spectrogram Images of Train Whistle Noise for various Noise Levels

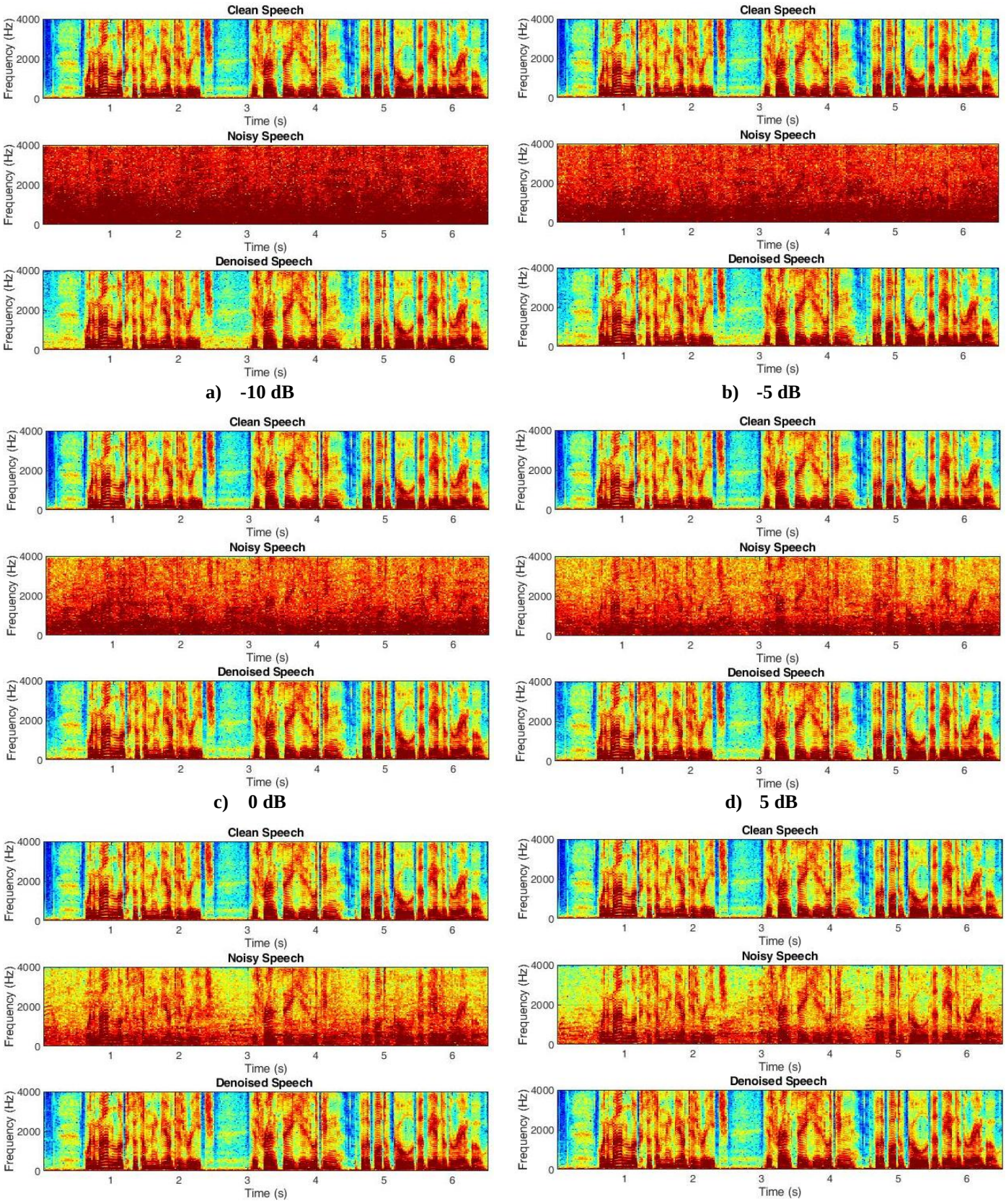
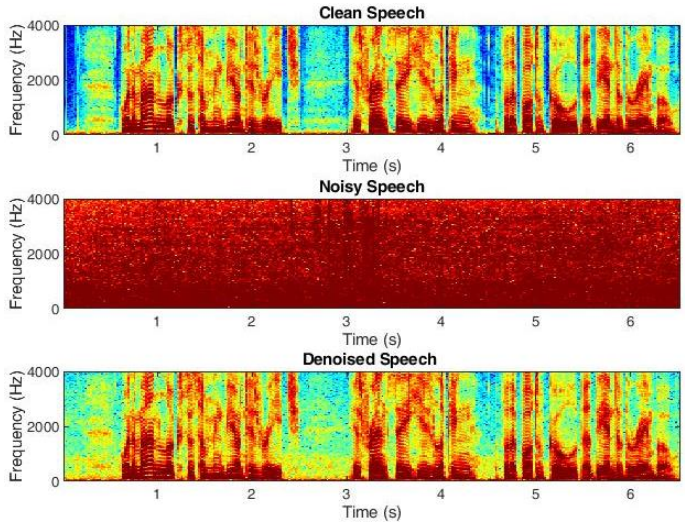
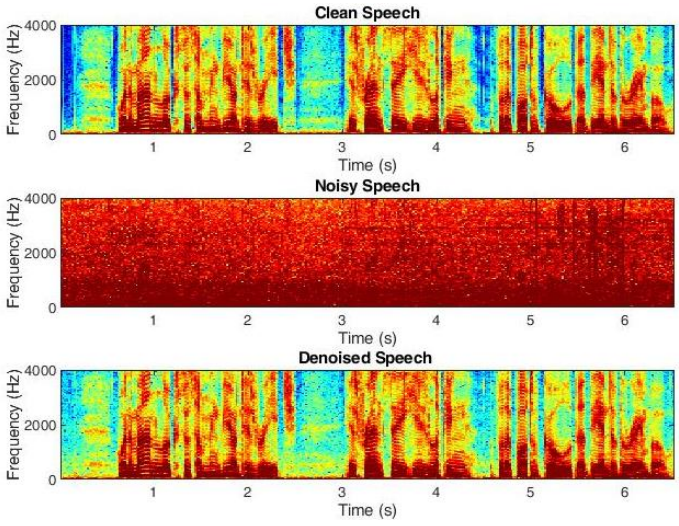


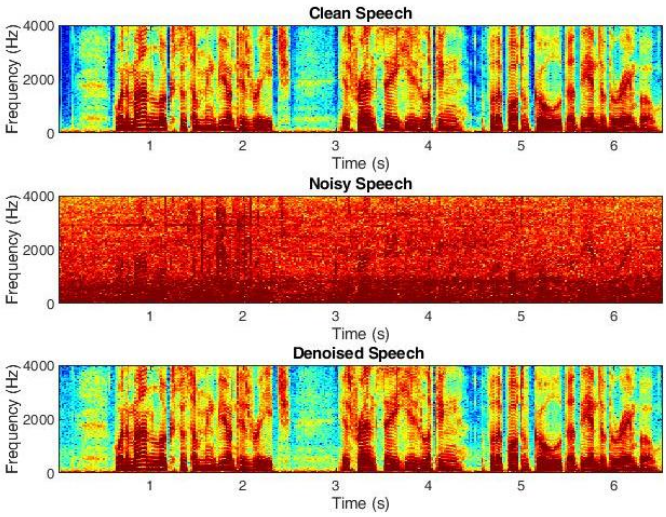
Figure 8 Deep CNN – Spectrogram Images of Restaurant Noise for various Noise Levels



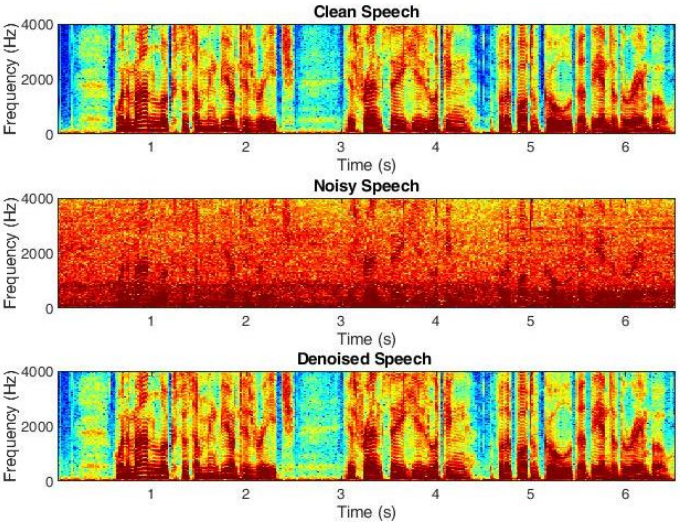
a) -10 dB



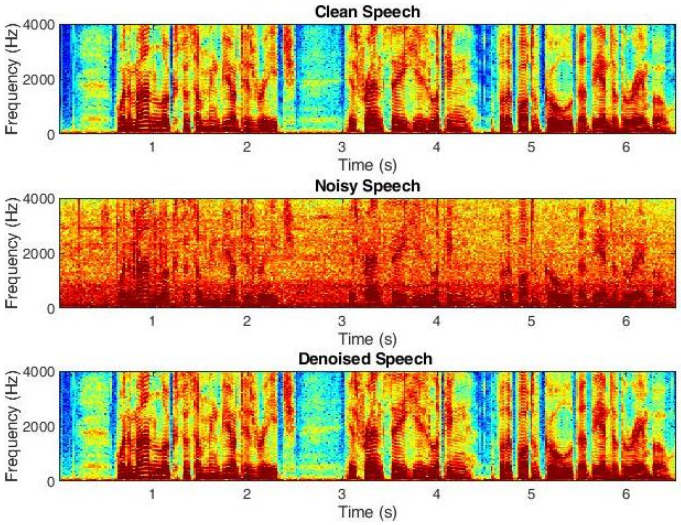
b) -5 dB



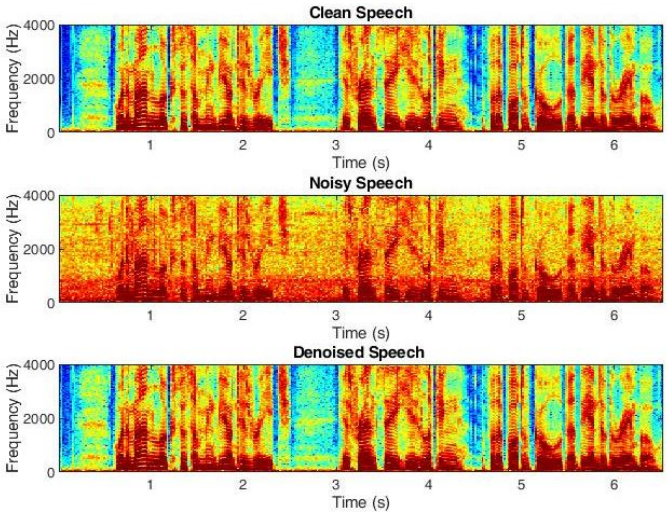
c) 0 dB



d) 5 dB

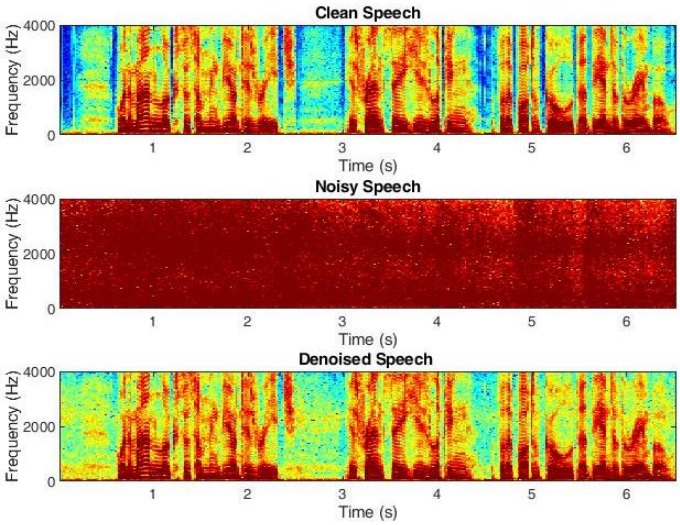


e) 10 dB

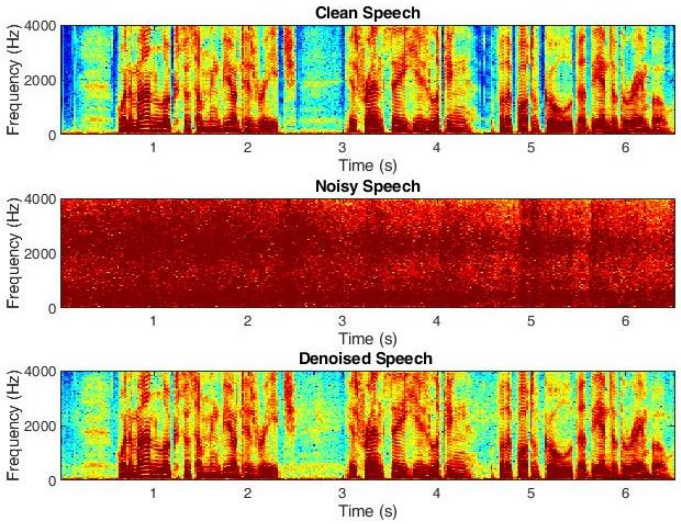


f) 15 dB

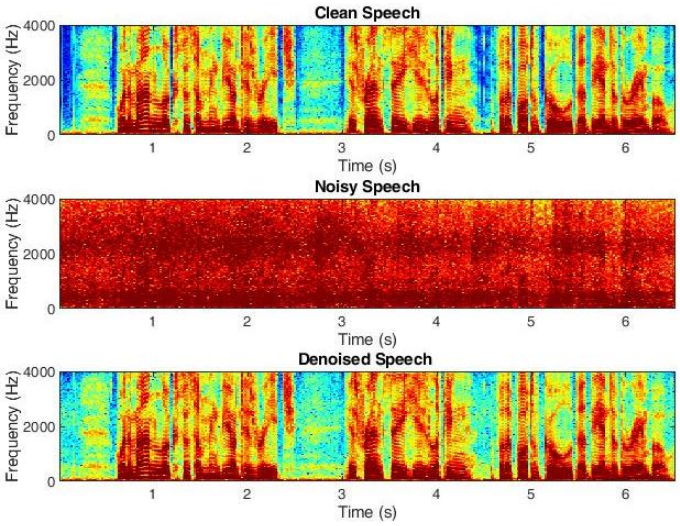
Figure 9 Deep CNN - Spectrogram Images of Car Noise for various Noise Level



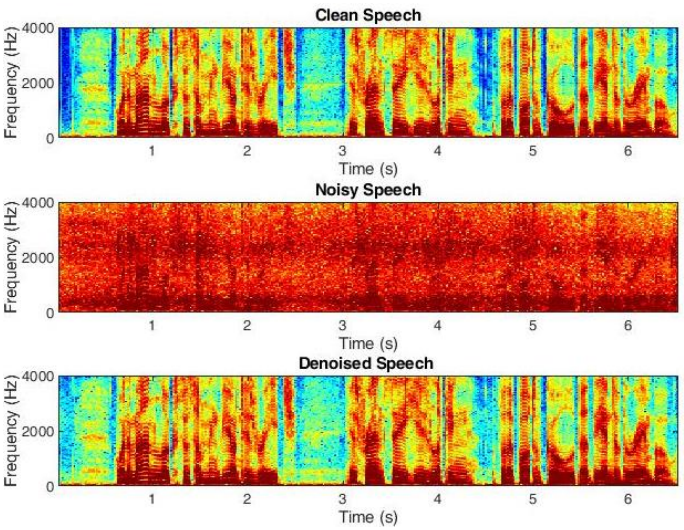
a) -10 dB



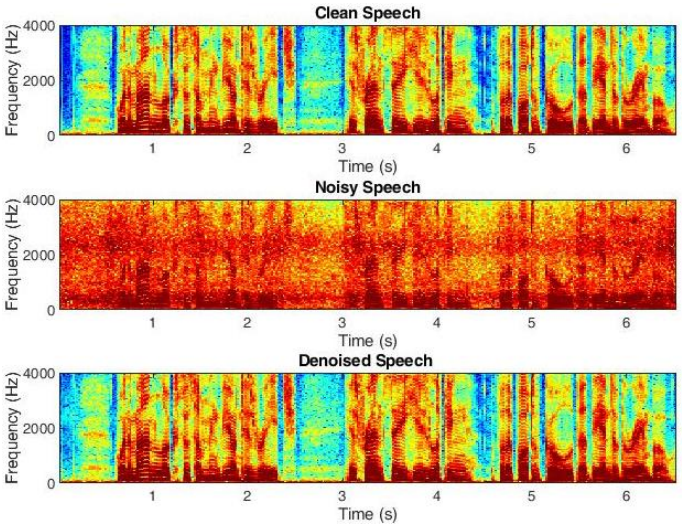
b) -5 dB



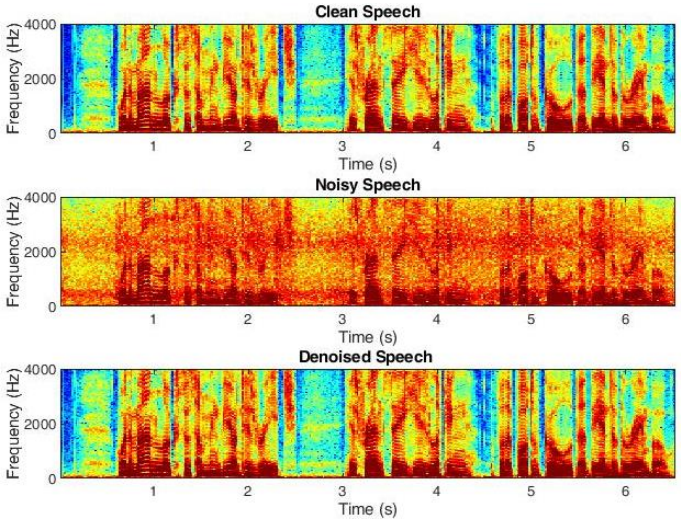
c) 0 dB



d) 5 dB



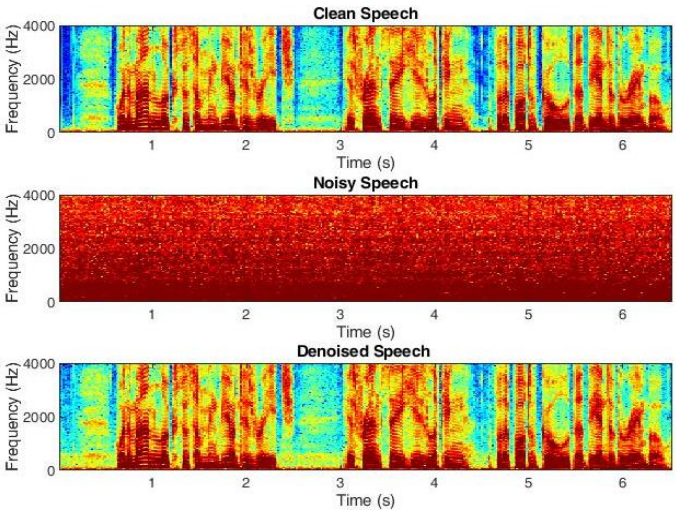
e) 10 dB



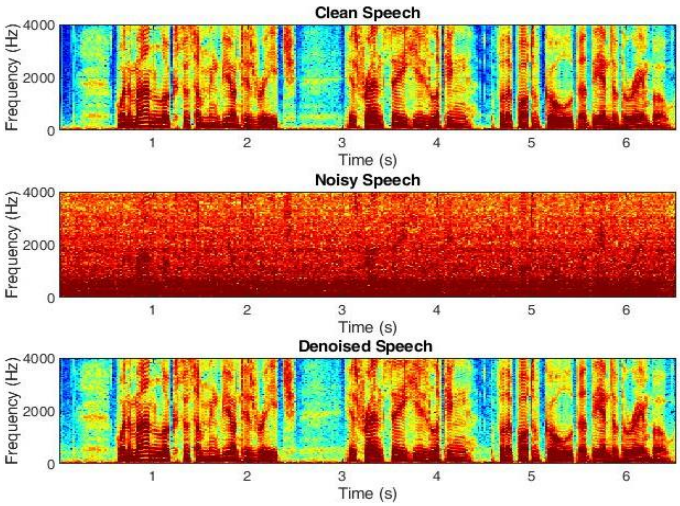
f) 15 dB

Figure 10 Deep CNN - Spectrogram Images of Subway Noise for various Noise Level

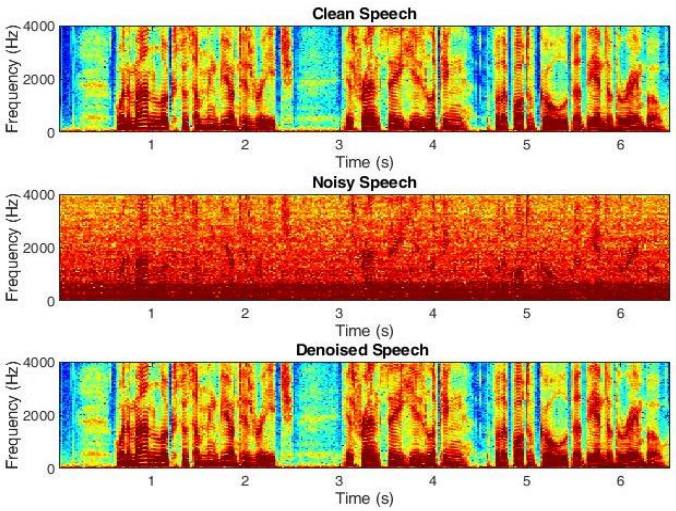
APPENDIX 3 (Spectrograms of Modified LSTM)



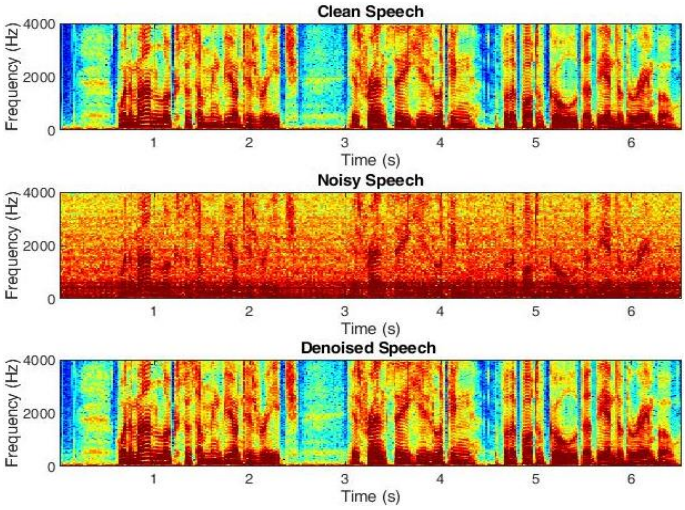
a) -10 dB



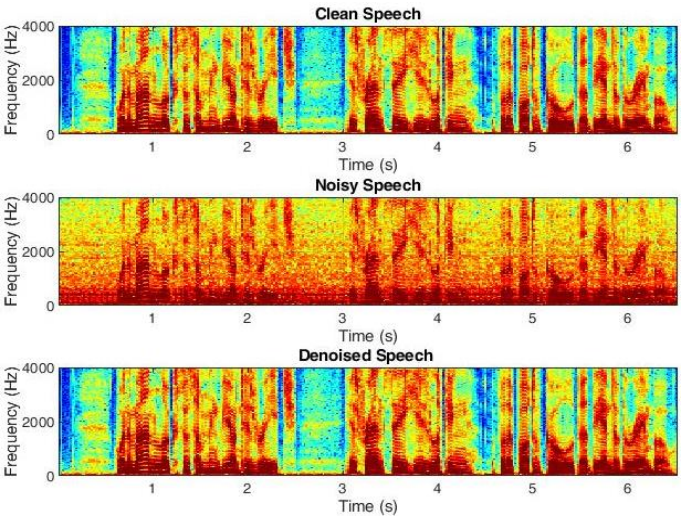
b) -5 dB



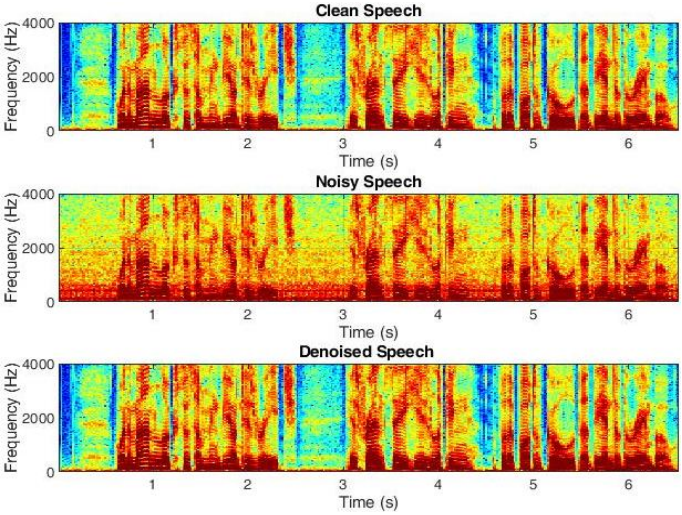
c) 0 dB



d) 5 dB



e) 10 dB



f) 15 dB

Figure 1 Modified LSTM - Spectrogram Images of Washing Machine Noise for various Noise Level

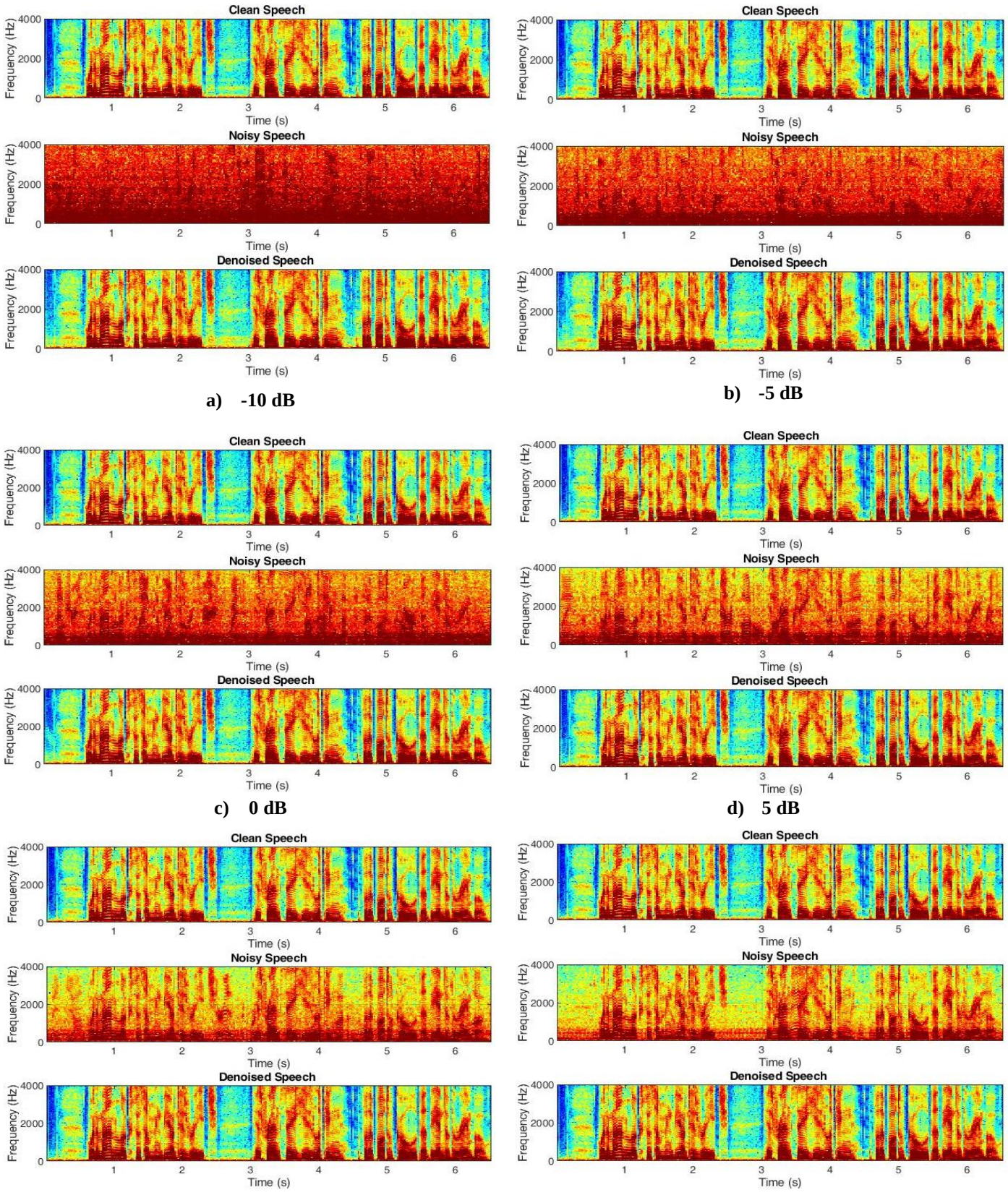
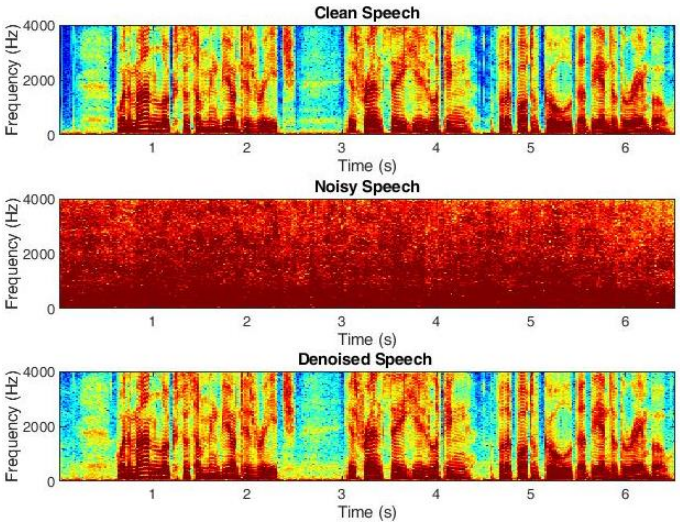
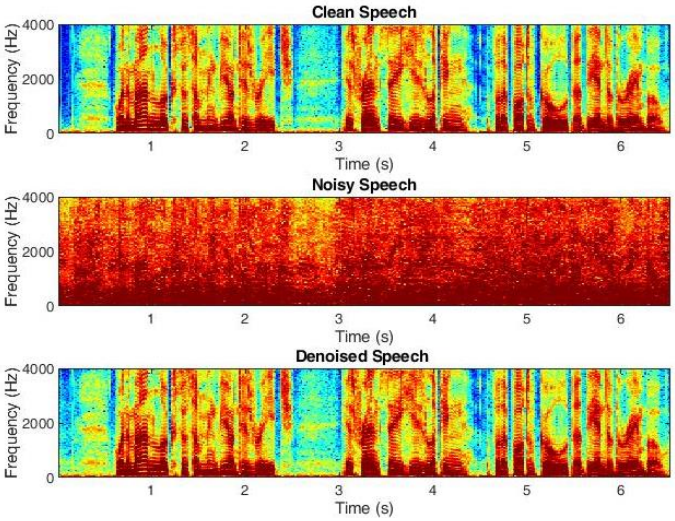


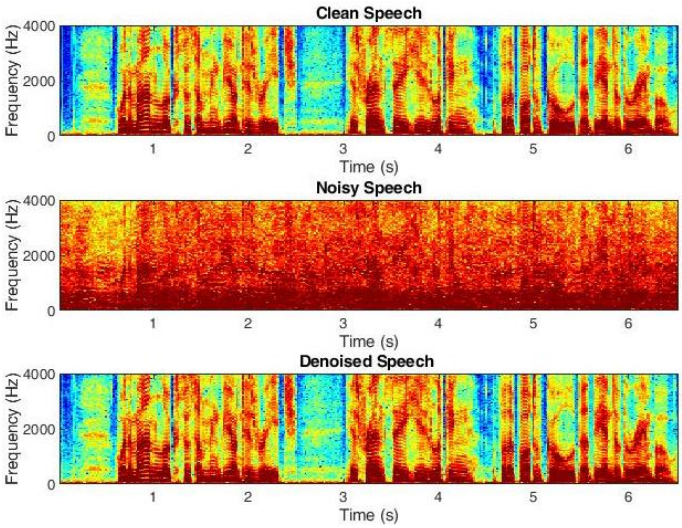
Figure 2 Modified LSTM - Spectrogram Images of Rainbow Noise for various Noise Levels



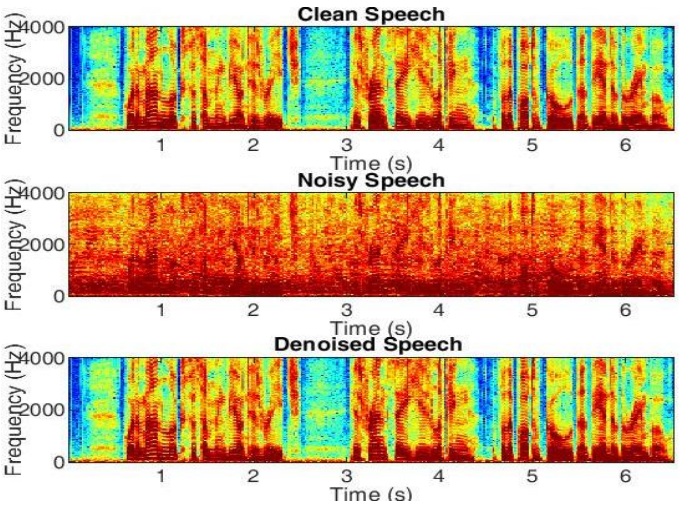
a) -10 dB



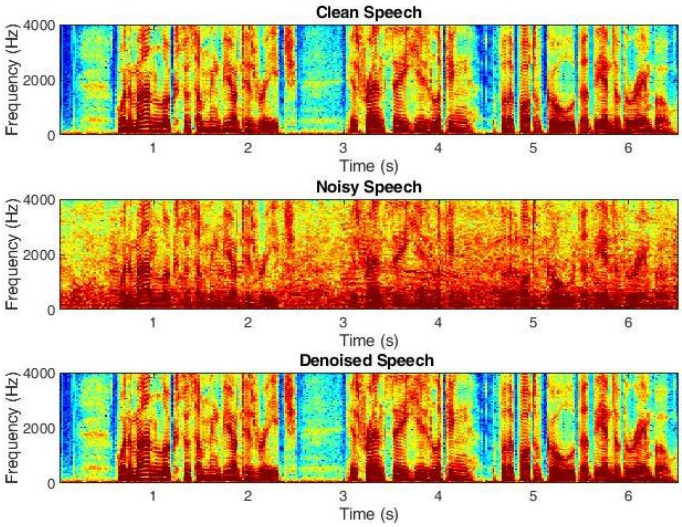
b) -5 dB



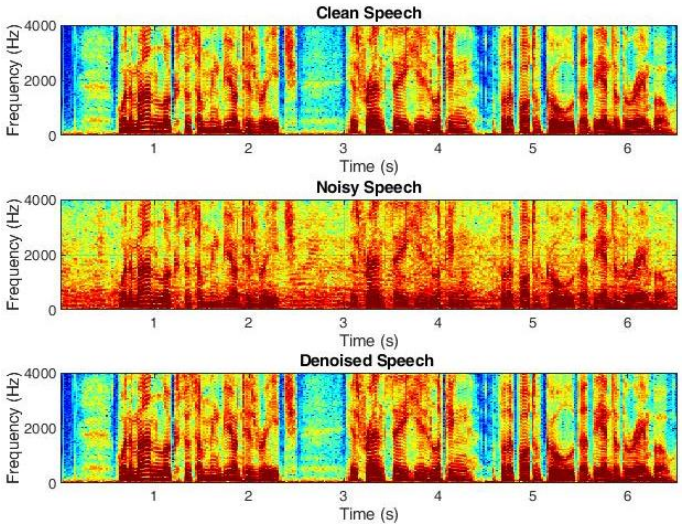
c) 0 dB



d) 5 dB



e) 10 dB



f) 15 dB

Figure 3 Modified LSTM - Spectrogram Images of Babble Noise for various Noise Levels

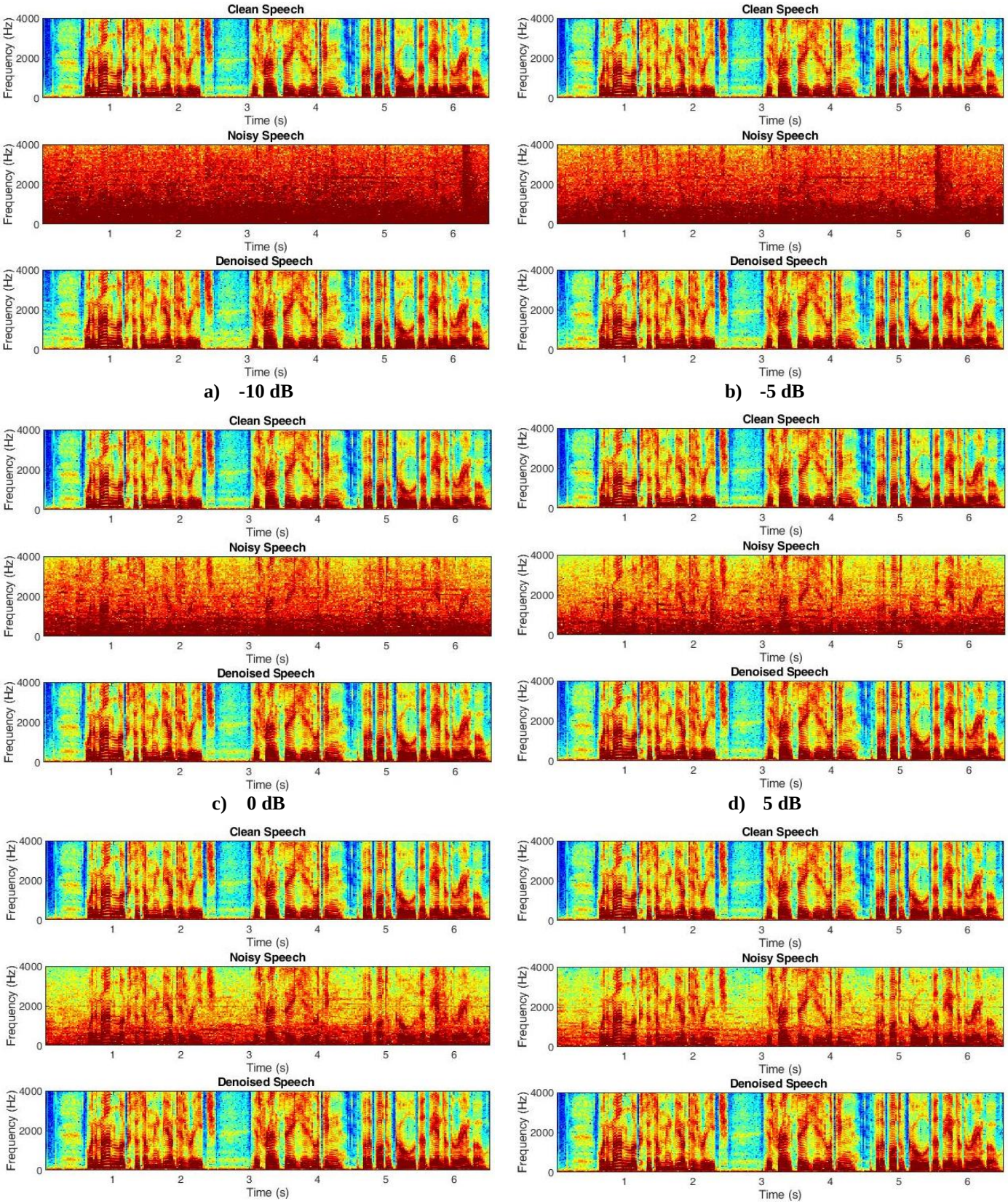
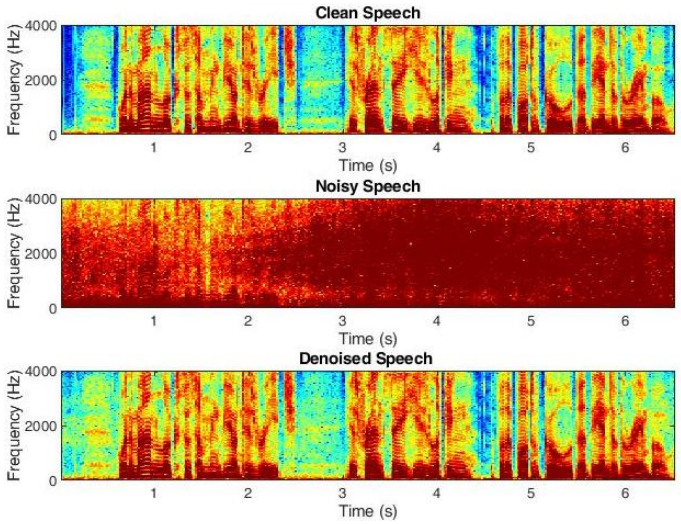
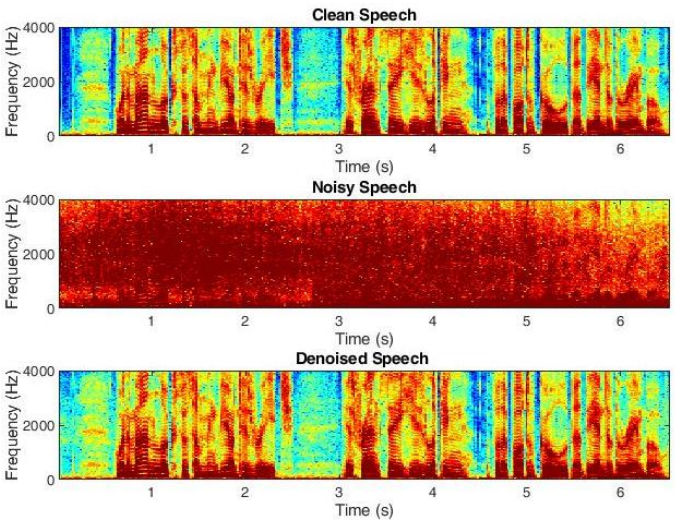


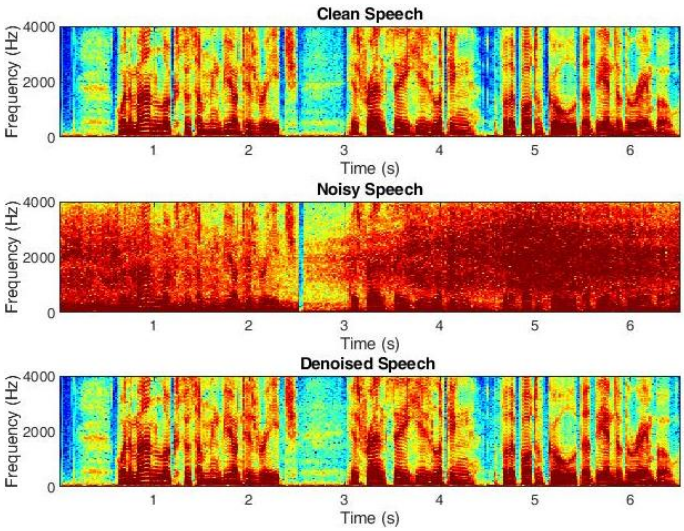
Figure 4 Modified LSTM - Spectrogram Images of Airport Noise for various Noise Levels



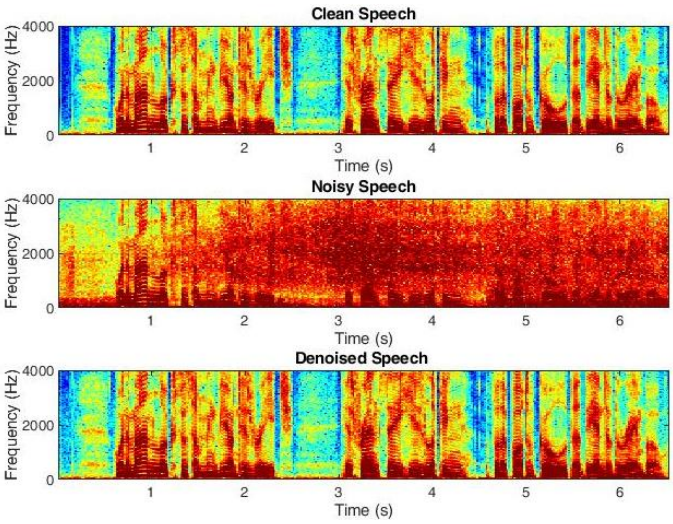
a) -10 dB



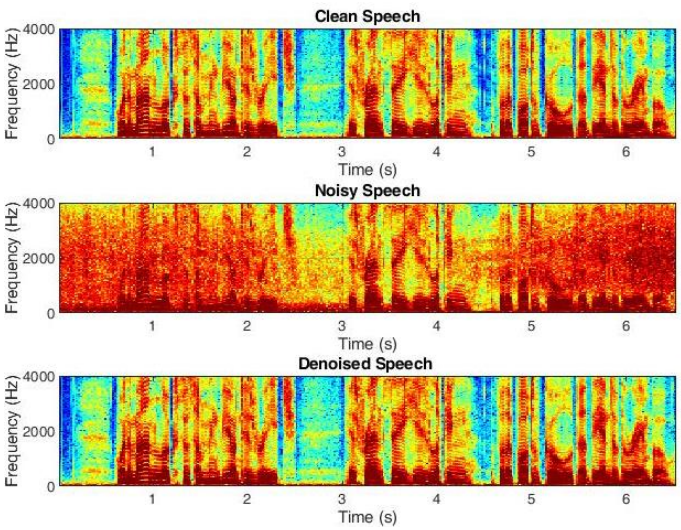
b) -5 dB



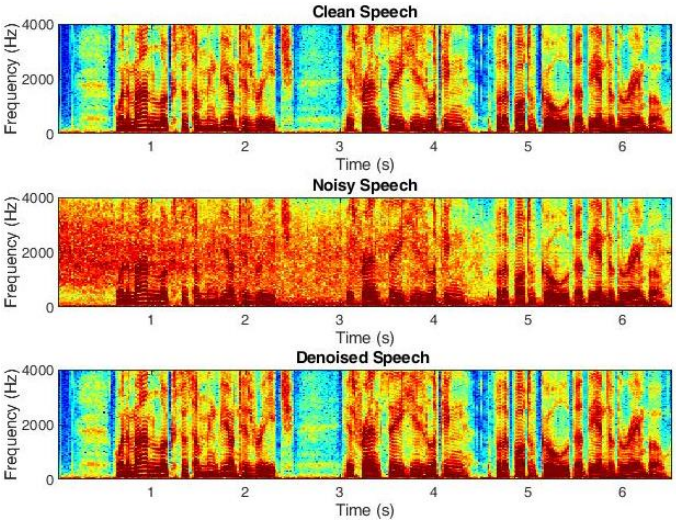
c) 0 dB



d) 5 dB

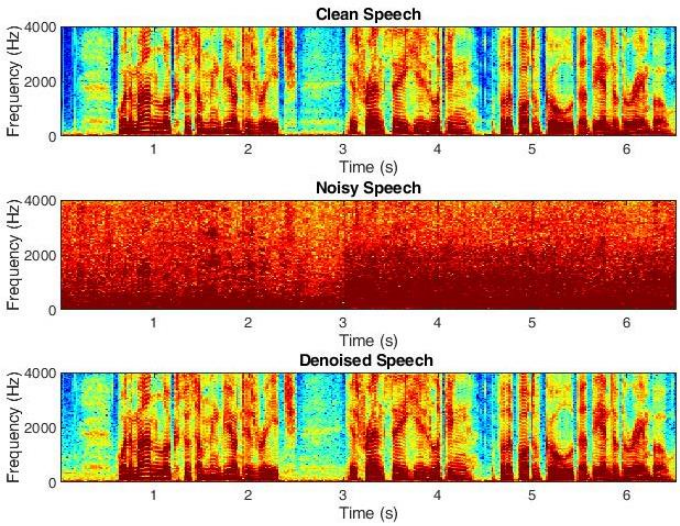


e) 10 dB

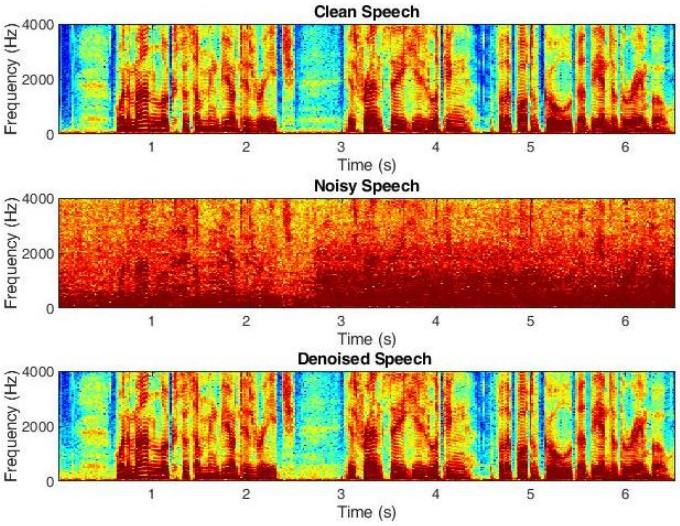


f) 15 dB

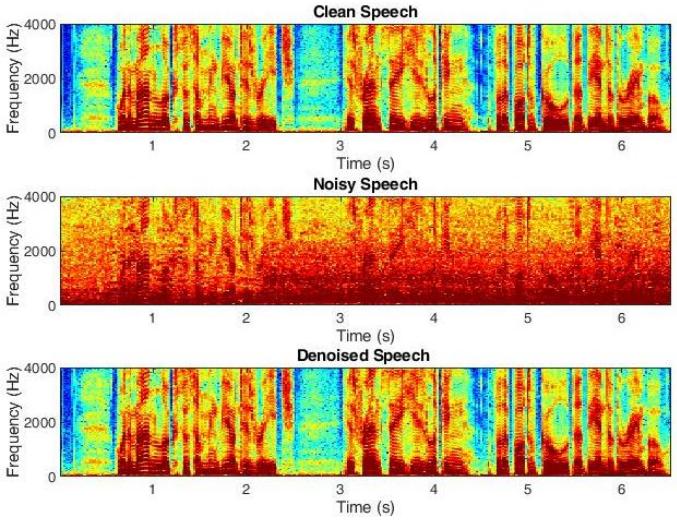
Figure 5 Modified LSTM - Spectrogram Images of Jetplane Noise for various Noise Levels



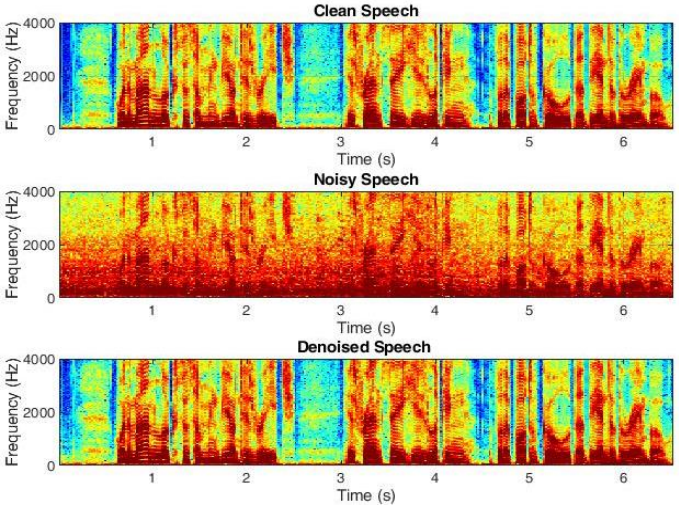
a) -10 dB



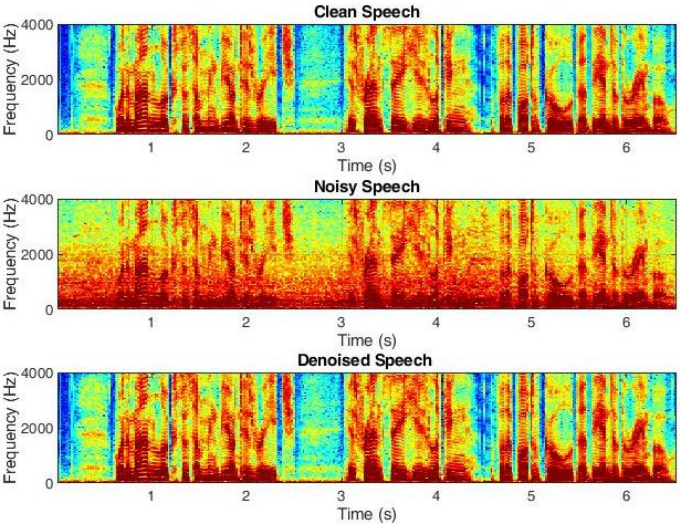
b) -5 dB



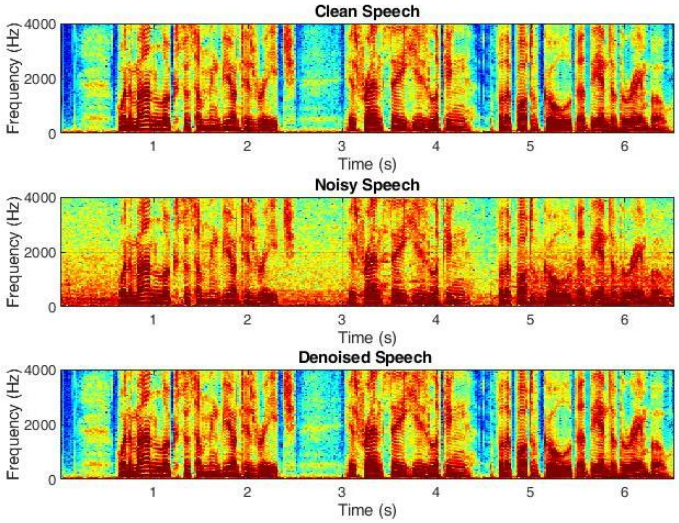
c) 0 dB



d) 5 dB

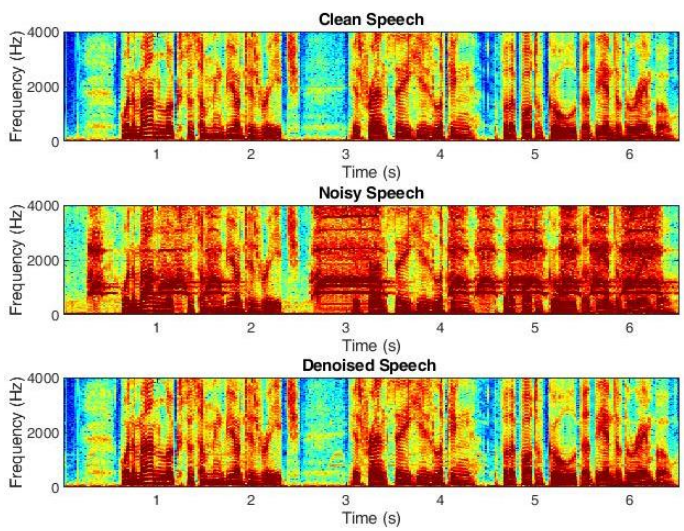


e) 10 dB

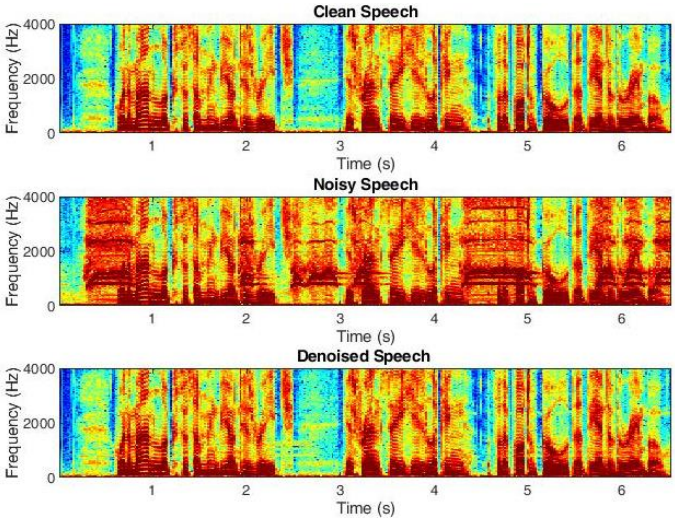


f) 15 dB

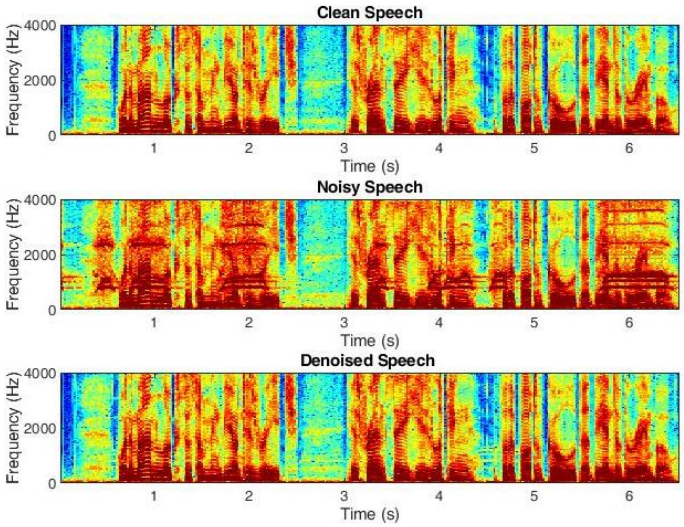
Figure 6 Modified LSTM - Spectrogram Images of Street Noise for various Noise Levels



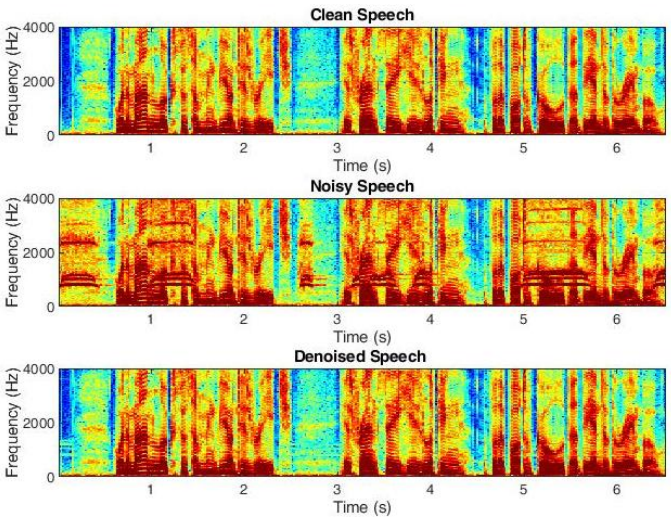
a) -10 dB



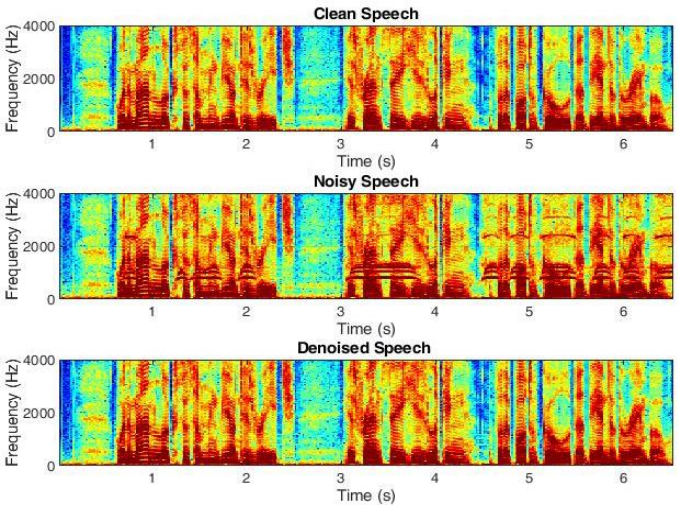
b) -5 dB



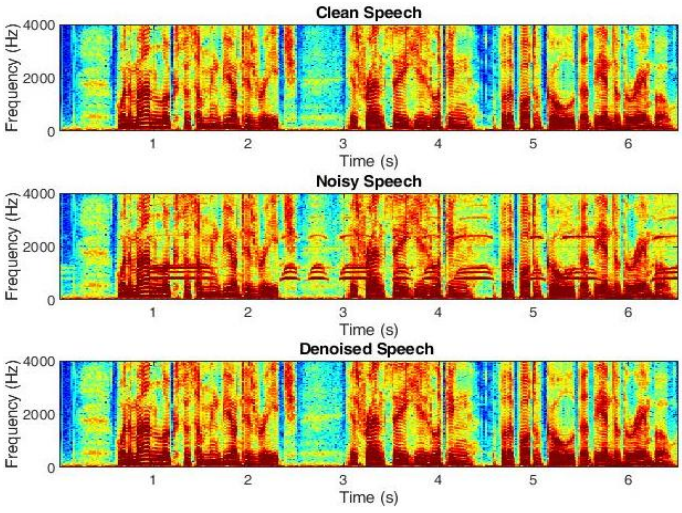
c) 0 dB



d) 5 dB

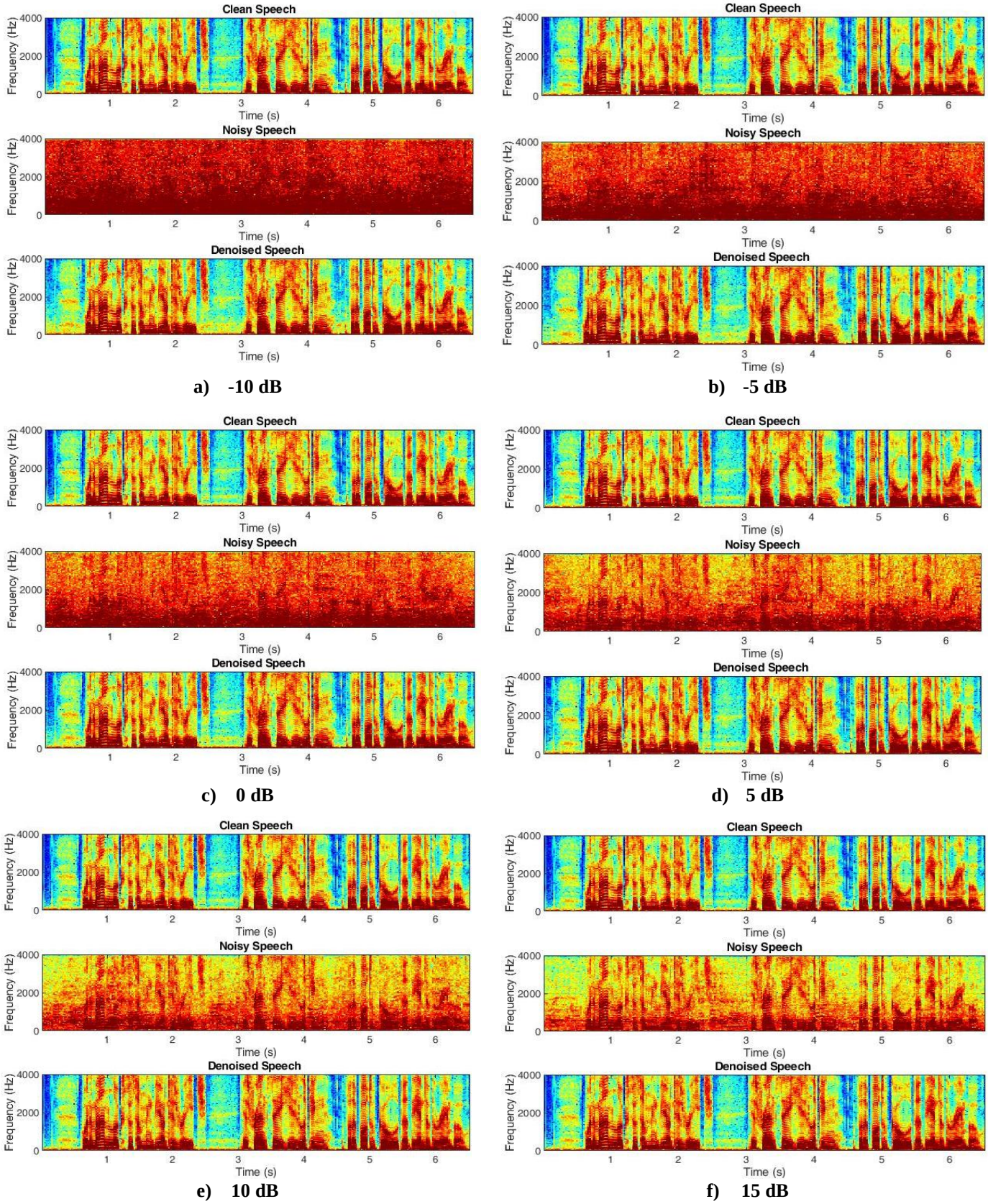


e) 10 dB



f) 15 dB

Figure 7 Modified LSTM - Spectrogram Images of Train Whistle Noise for various Noise Level

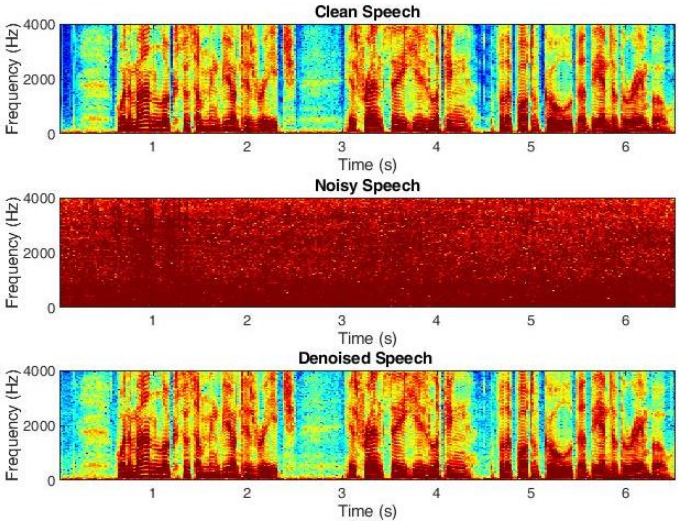


a) -10 dB b) -5 dB

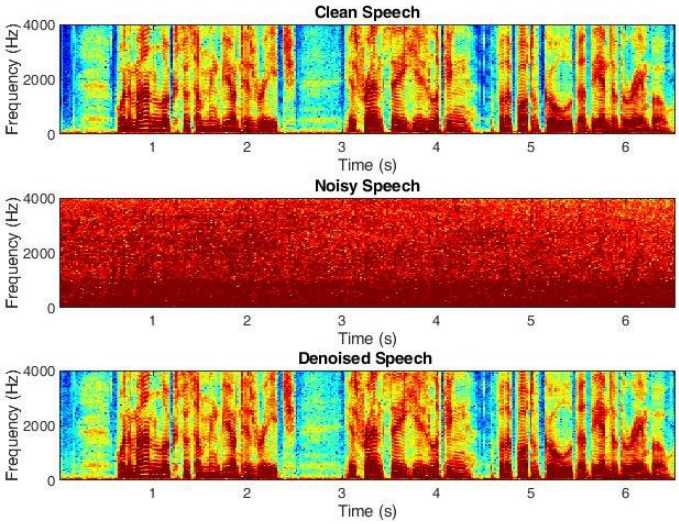
c) 0 dB d) 5 dB

e) 10 dB f) 15 dB

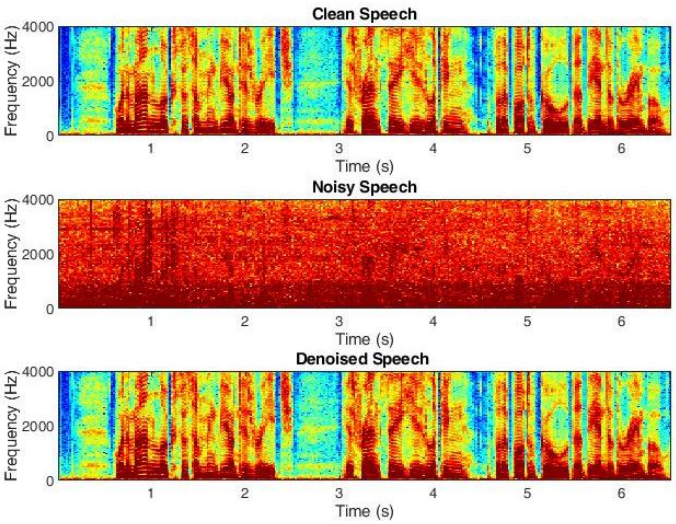
Figure 8 Modified LSTM - Spectrogram Images of Restaurant Noise for various Noise Levels



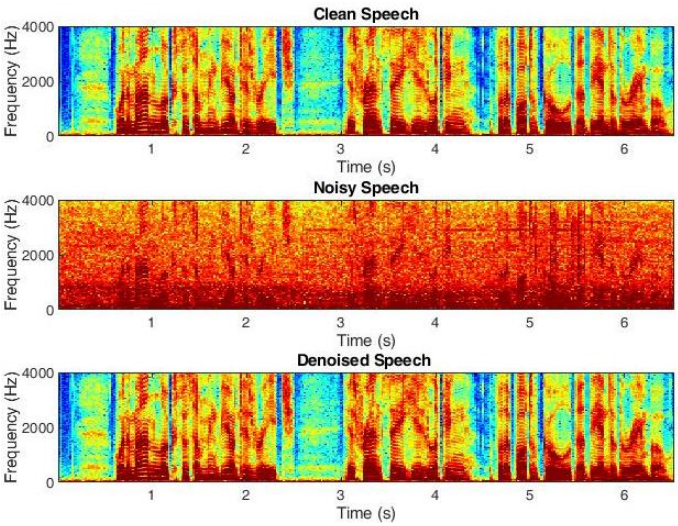
a) -10 dB



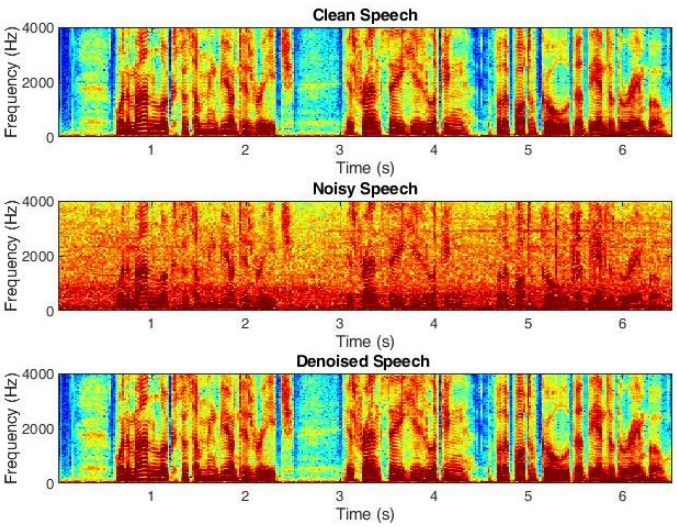
b) -5 dB



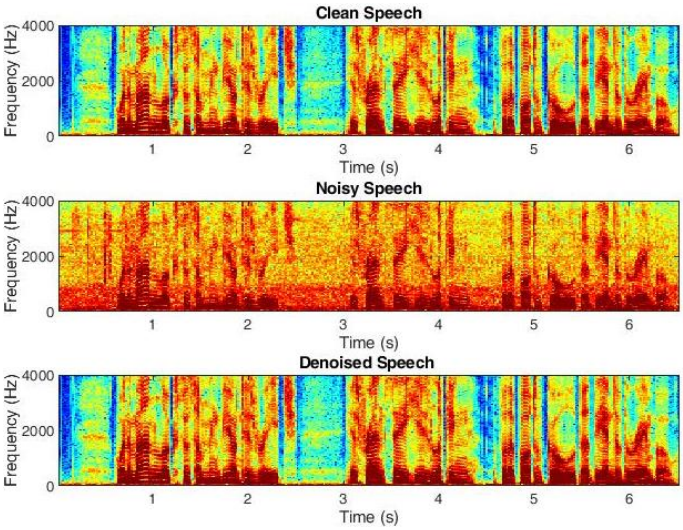
c) 0 dB



d) 5 dB

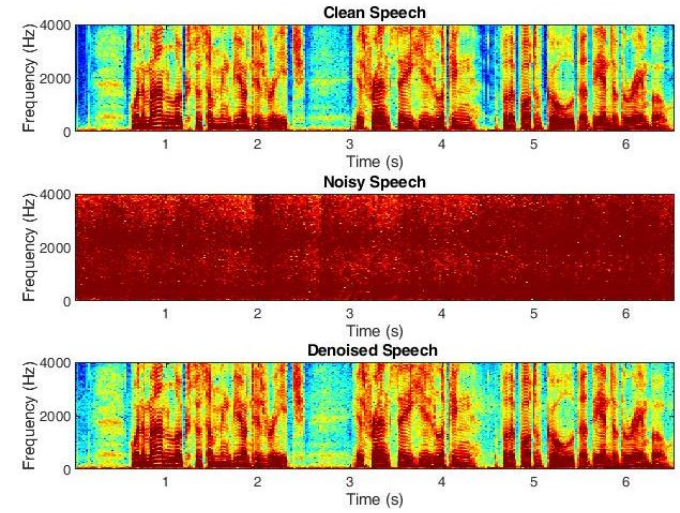


e) 10 dB

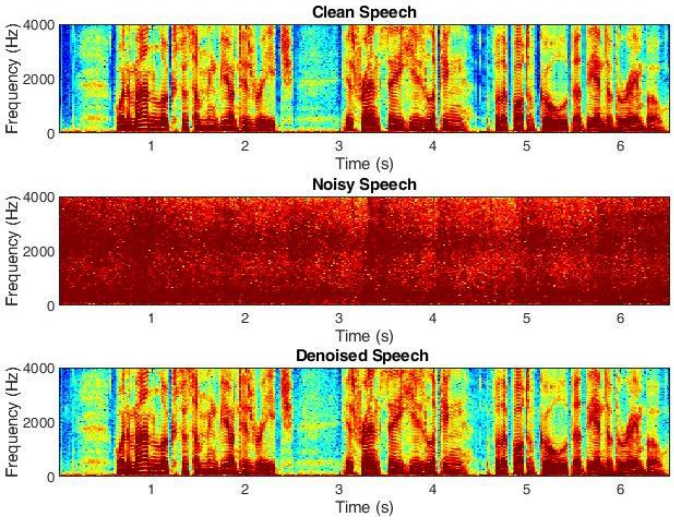


f) 15 dB

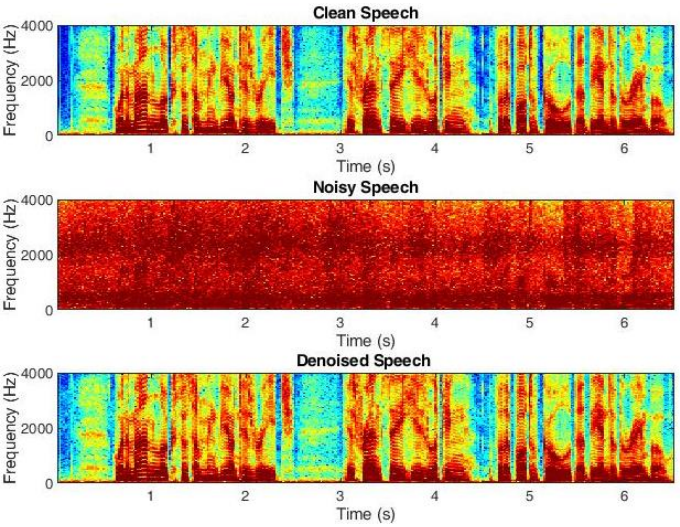
Figure 9 Modified LSTM - Spectrogram Images of Car Noise for various Noise Levels



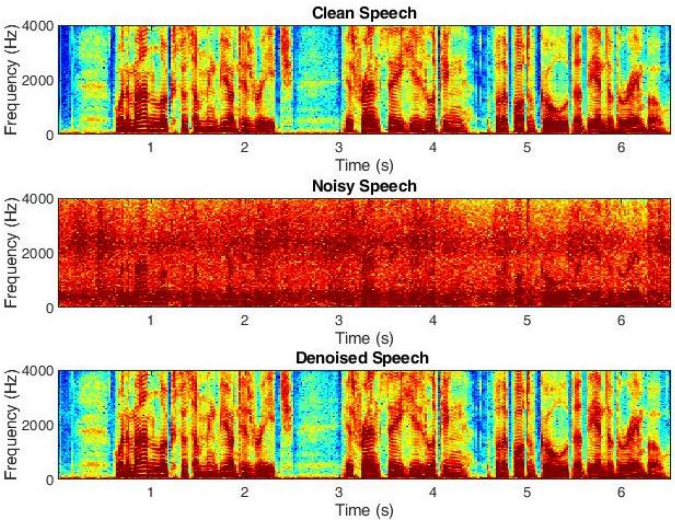
a) -10 dB



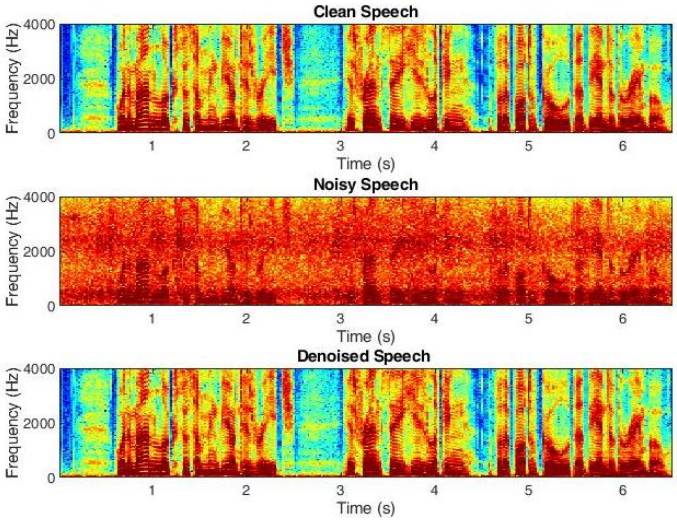
b) -5 dB



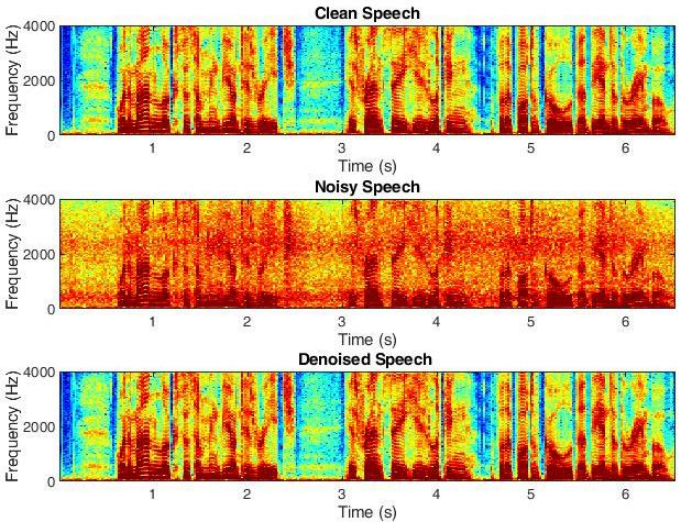
c) 0 dB



d) 5 dB



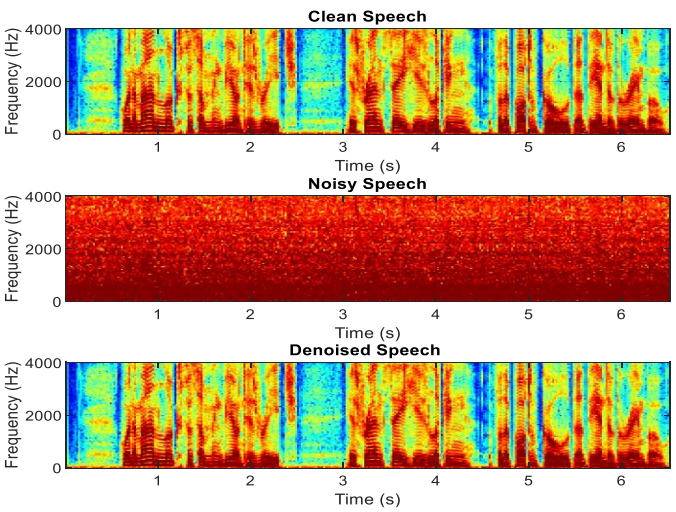
e) 10 dB



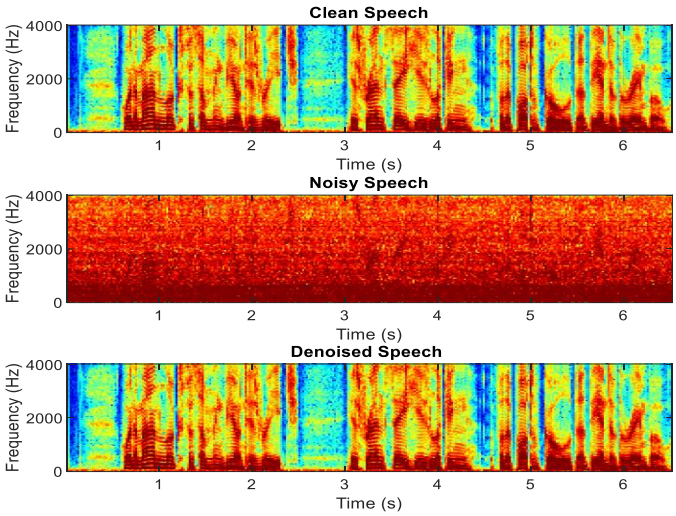
f) 15 dB

Figure 10 Modified LSTM - Spectrogram Images of Subway Noise for various Noise Levels

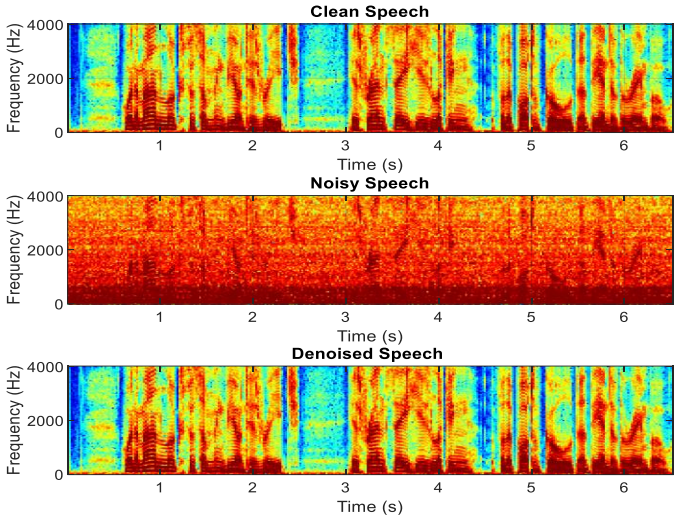
APPENDIX 4 (Spectrograms of Modified FCRN)



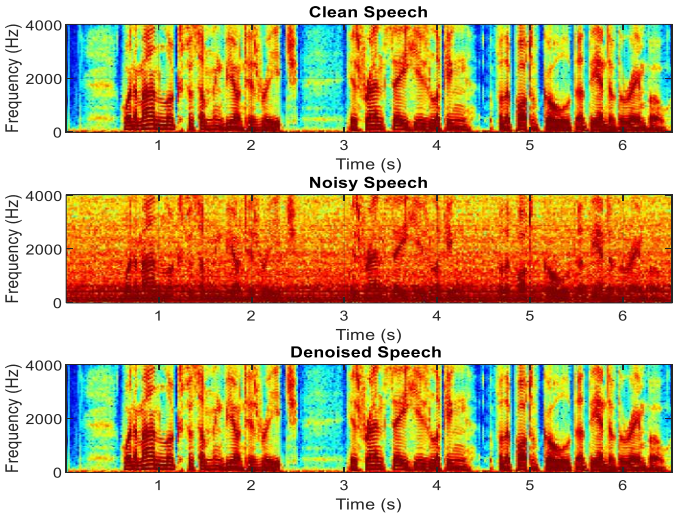
a) -10 dB



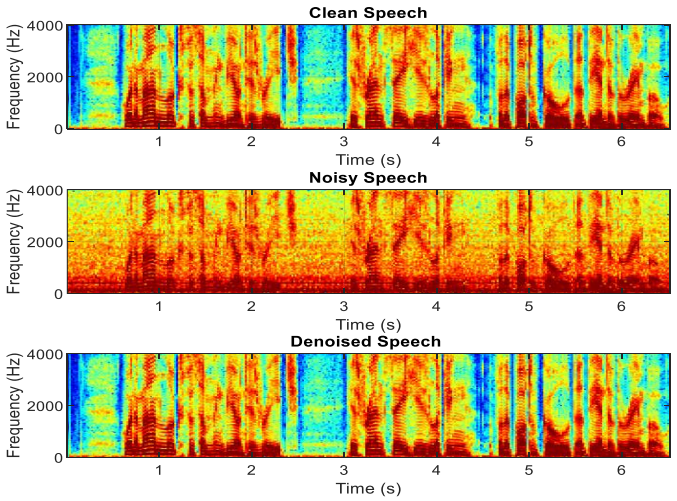
b) -5 dB



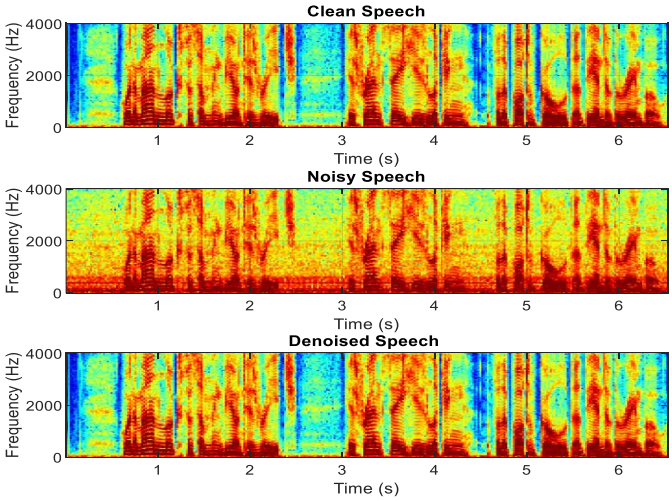
c) 0 dB



d) 5 dB



e) 10 dB



f) 15 dB

Figure 1 Modified FCRN - Spectrogram Images of Washing Machine Noise for various Noise Levels

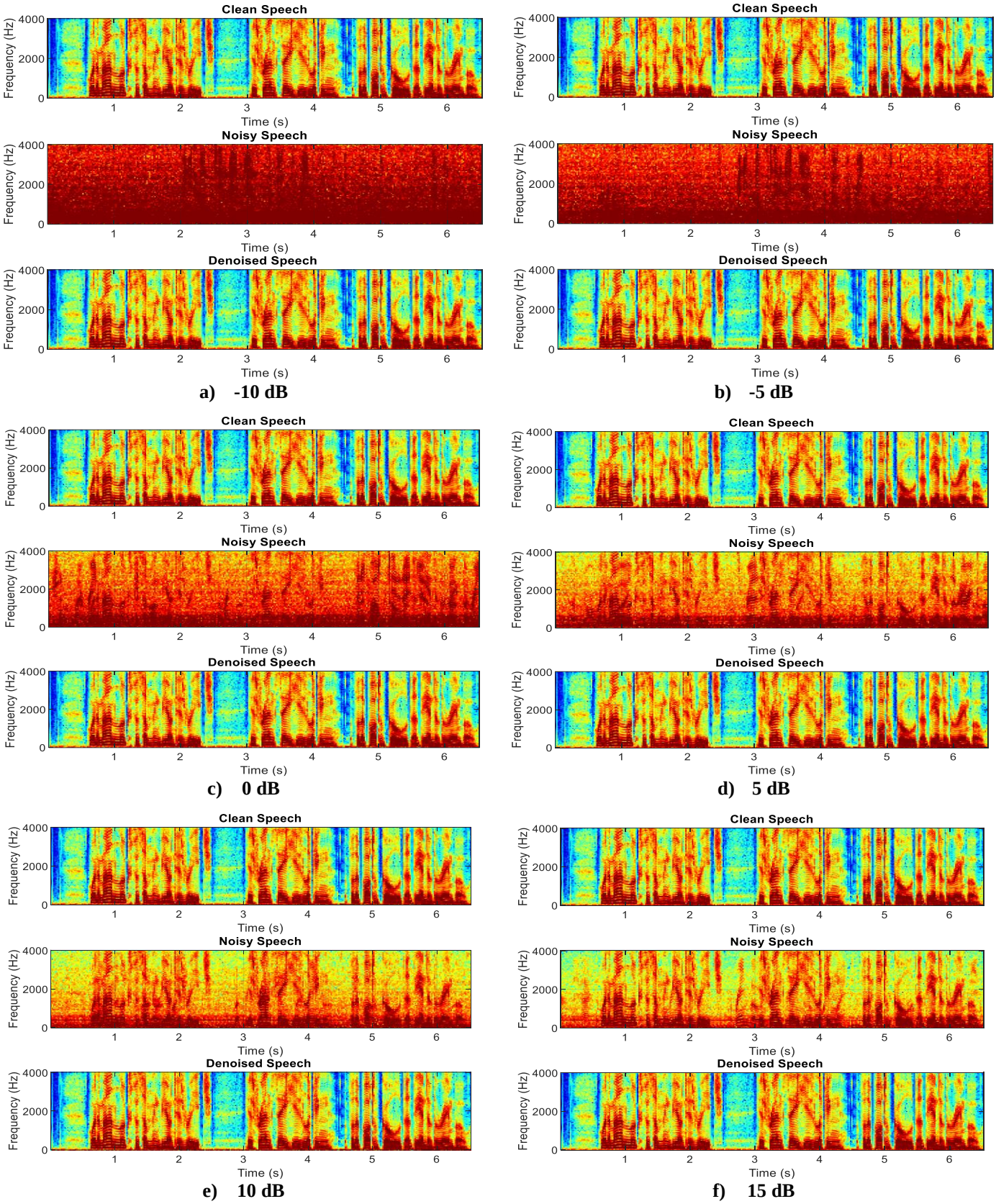
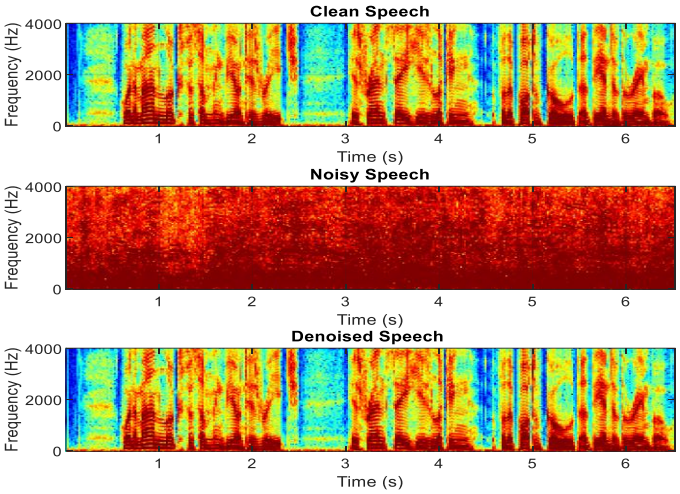
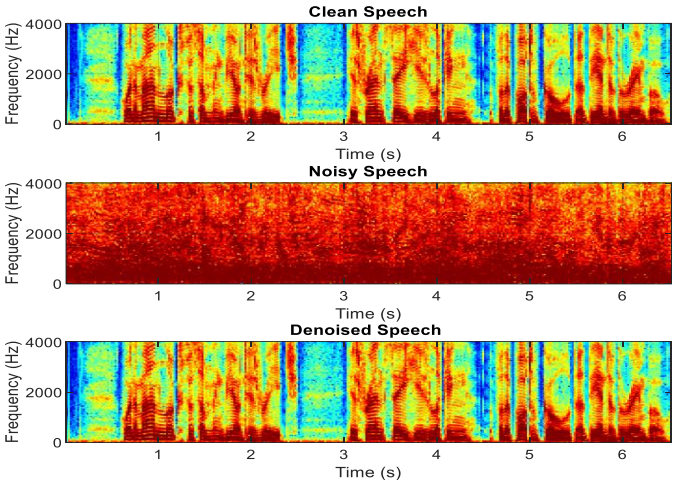


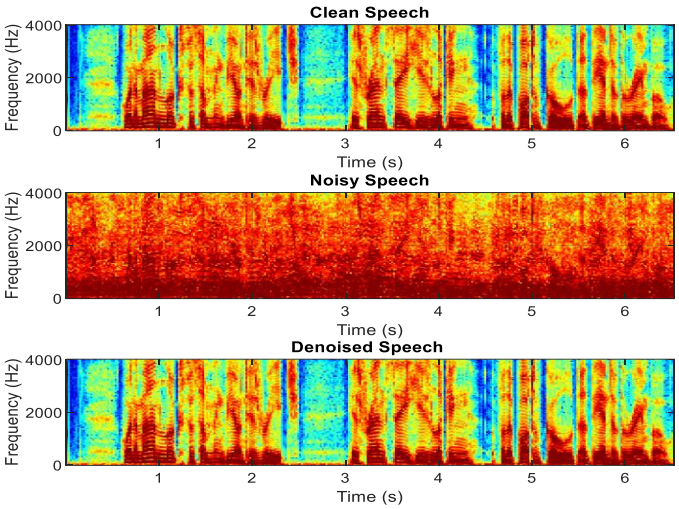
Figure 2 Modified FCRN - Spectrogram Images of Rainbow Noise for various Noise Levels



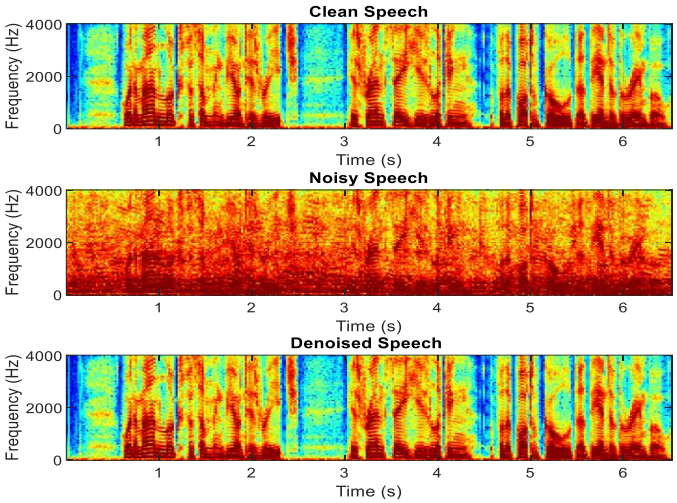
g) -10 dB



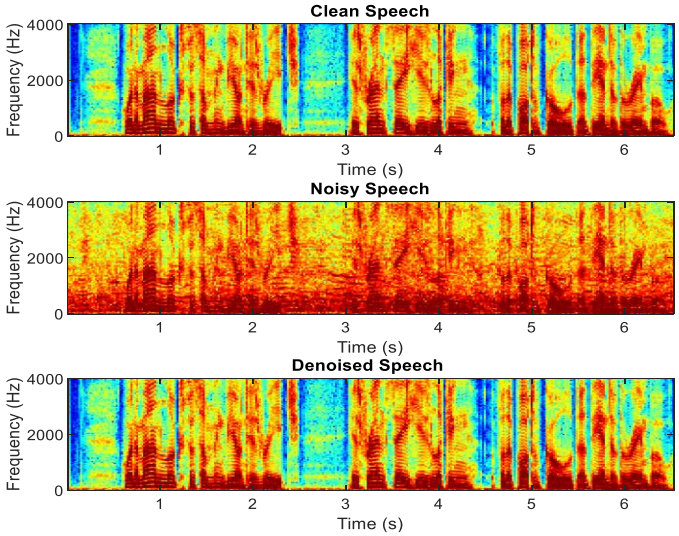
h) -5 dB



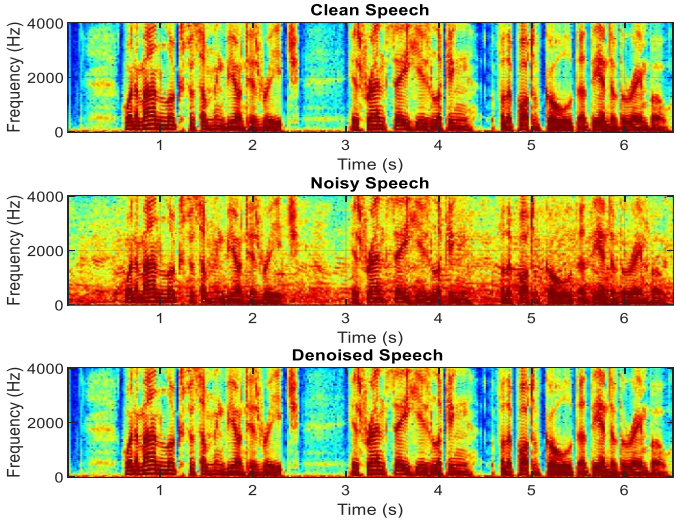
i) 0 dB



j) 5 dB

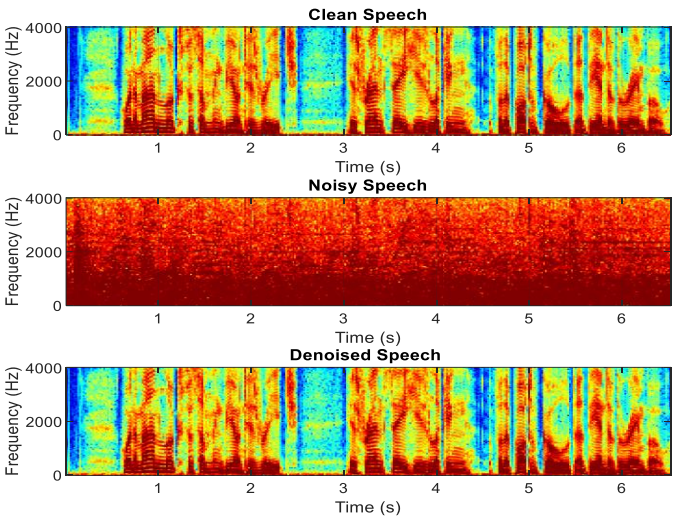


k) 10 dB

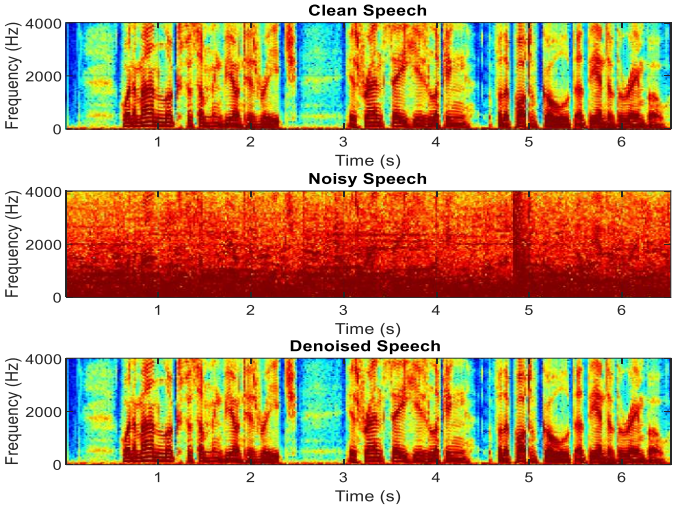


l) 15 dB

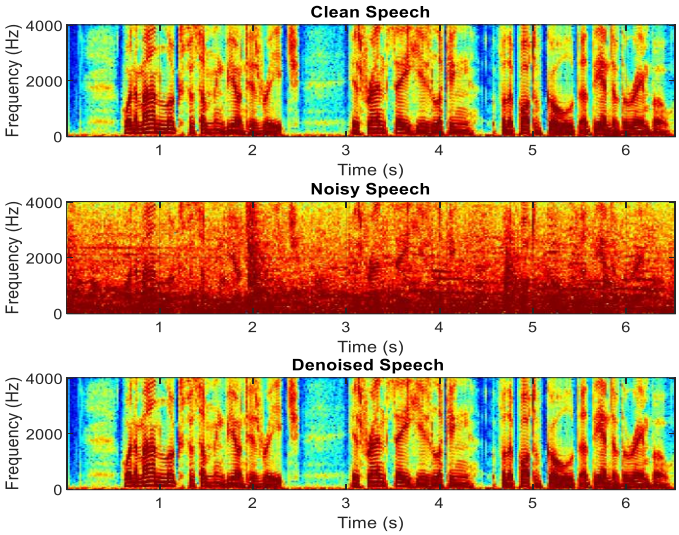
Figure 3 Modified FCRN - Spectrogram Images of Babble Noise for various Noise Levels



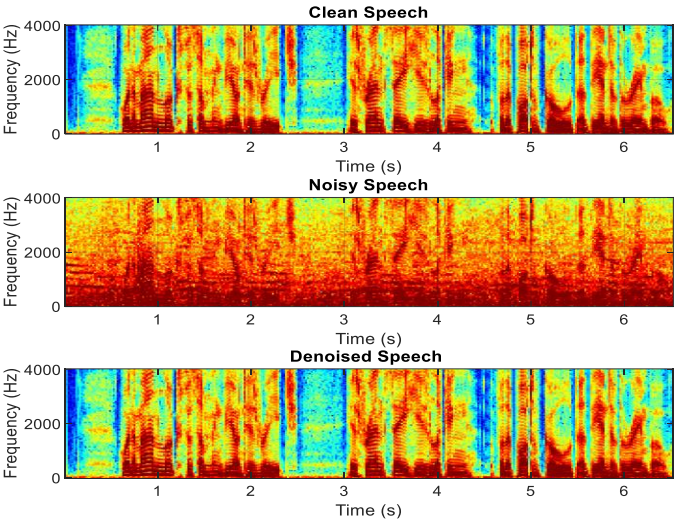
a) -10 dB



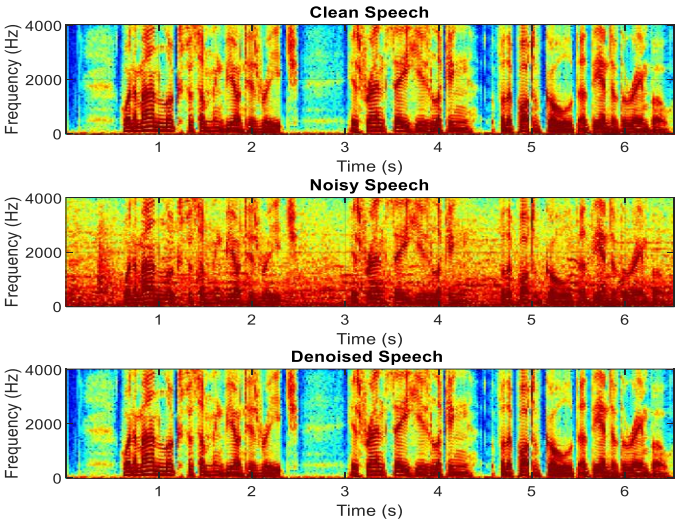
b) -5 dB



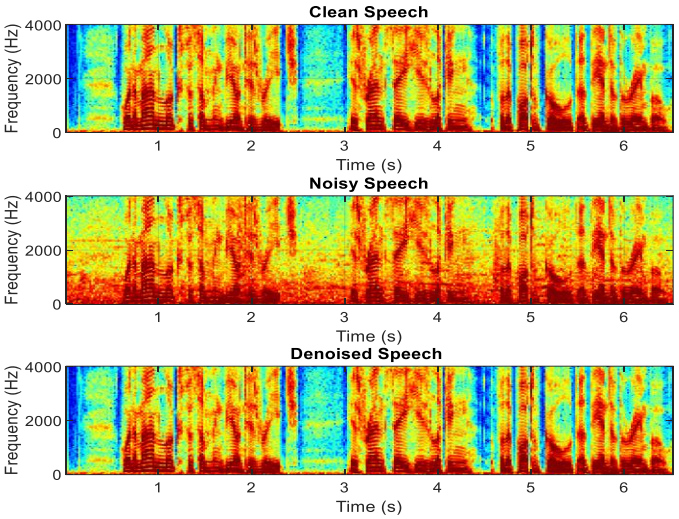
c) 0 dB



d) 5 dB



e) 10 dB



f) 15 dB

Figure 4 Modified FCRN - Spectrogram Images of Airport Noise for various Noise Levels

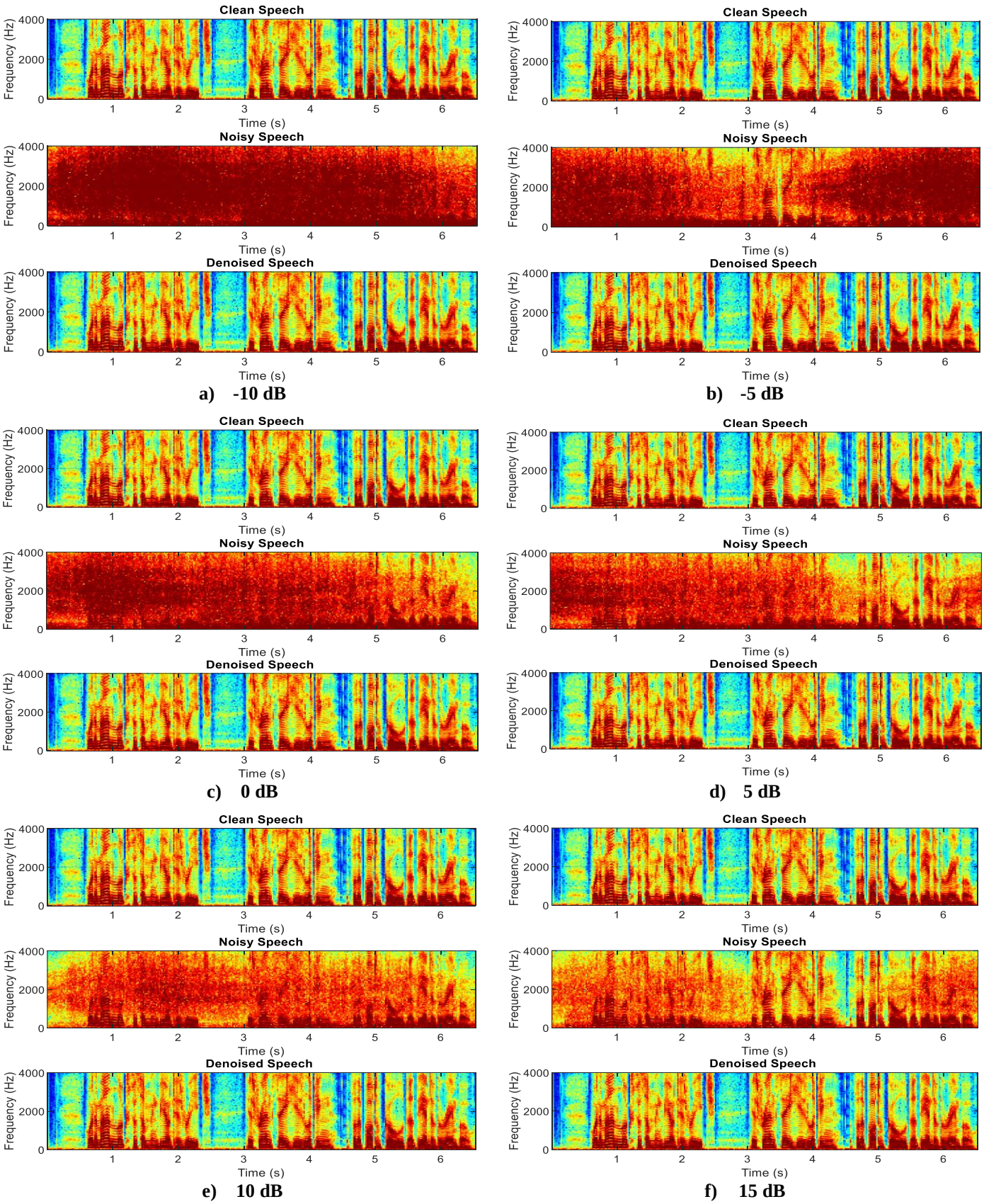
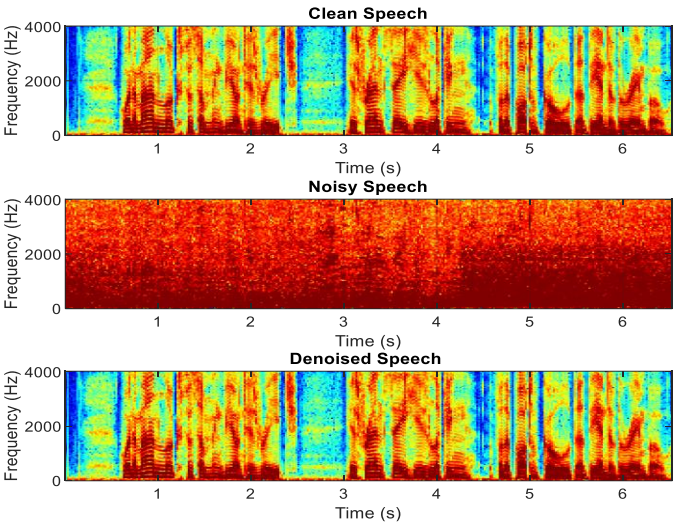
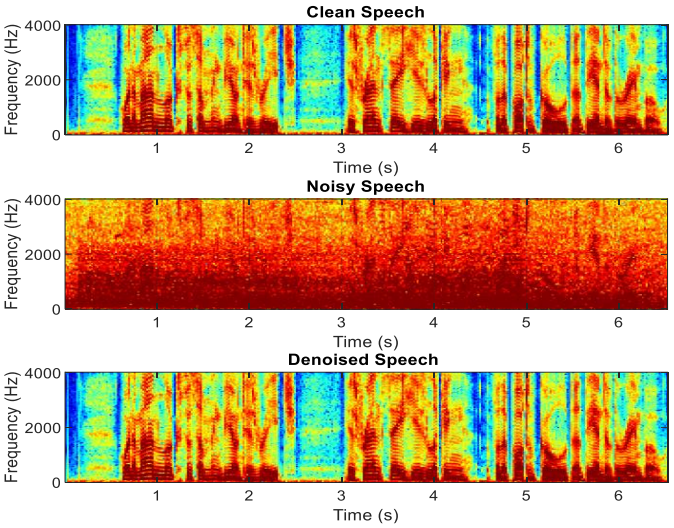


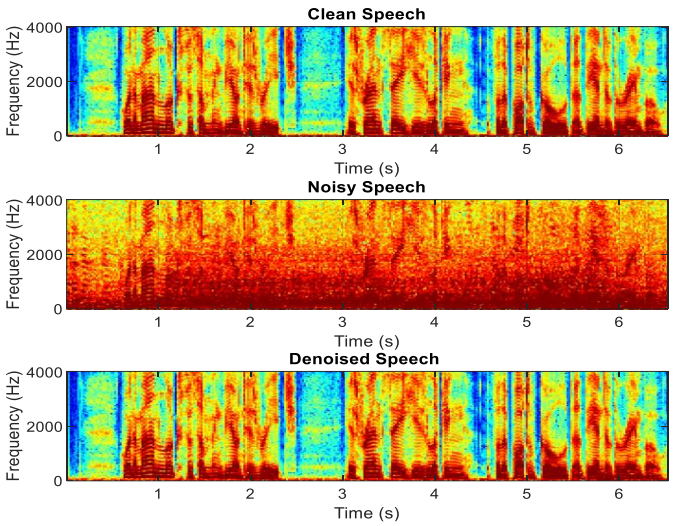
Figure 5 Modified FCRN - Spectrogram Images of Jetplane Noise for various Noise Levels



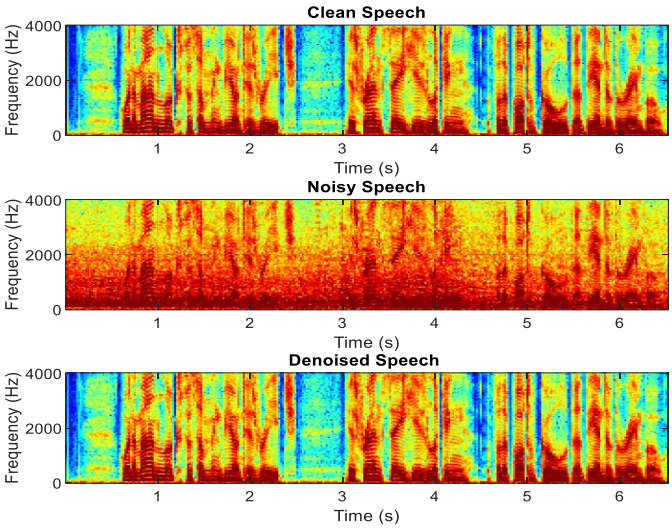
a) -10 dB



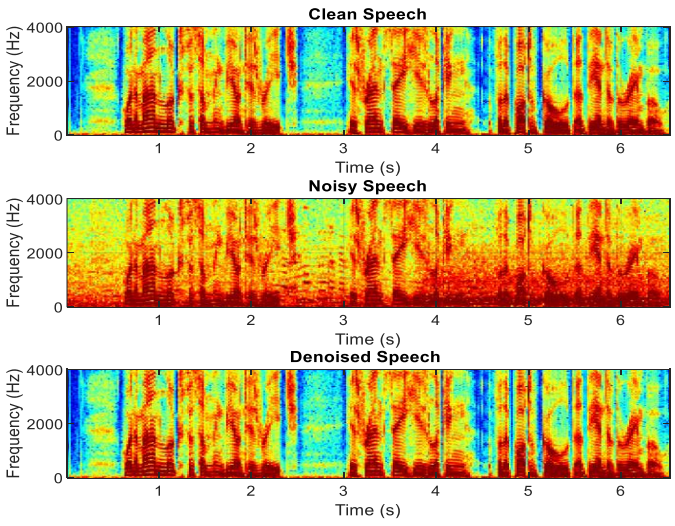
b) -5 dB



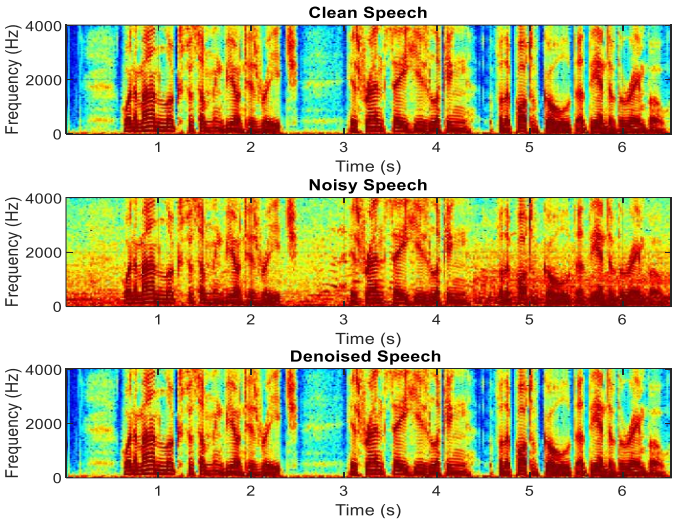
c) 0 dB



d) 5 dB

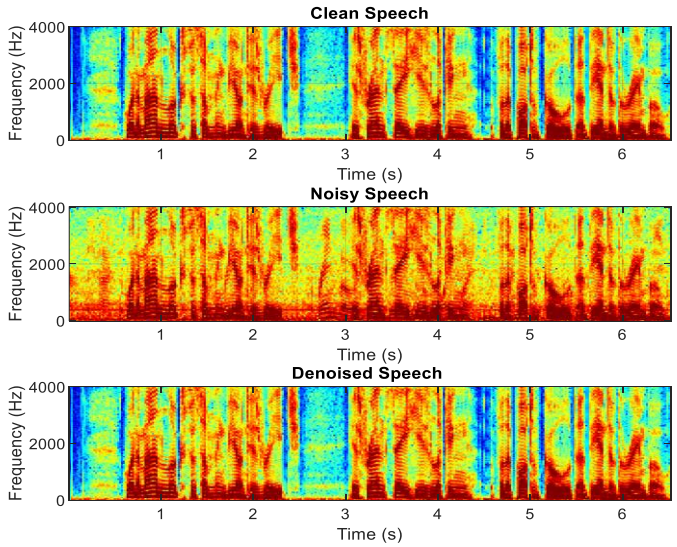


e) 10 dB

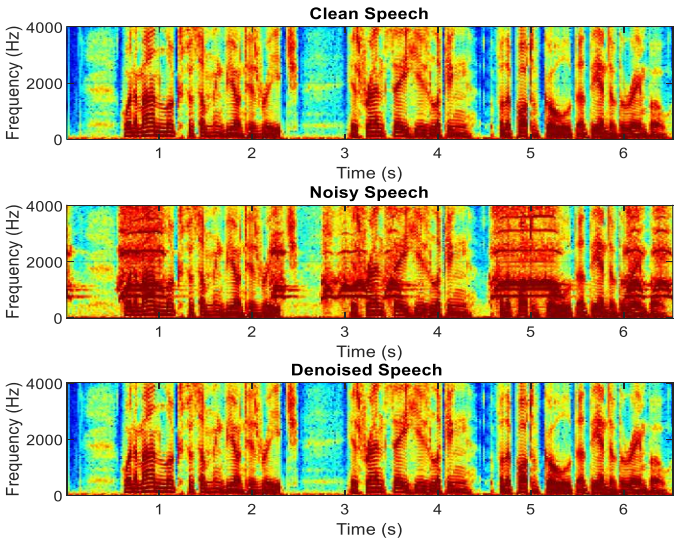


f) 15 dB

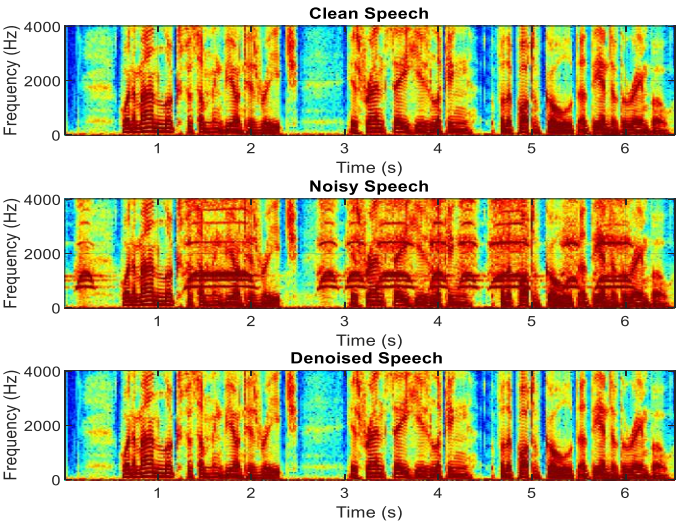
Figure 6 Modified FCRN - Spectrogram Images of Street Noise for various Noise Levels



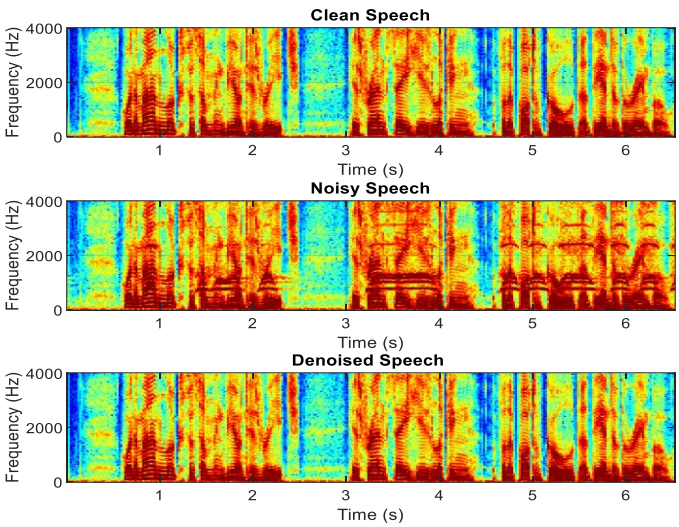
g) -10 dB



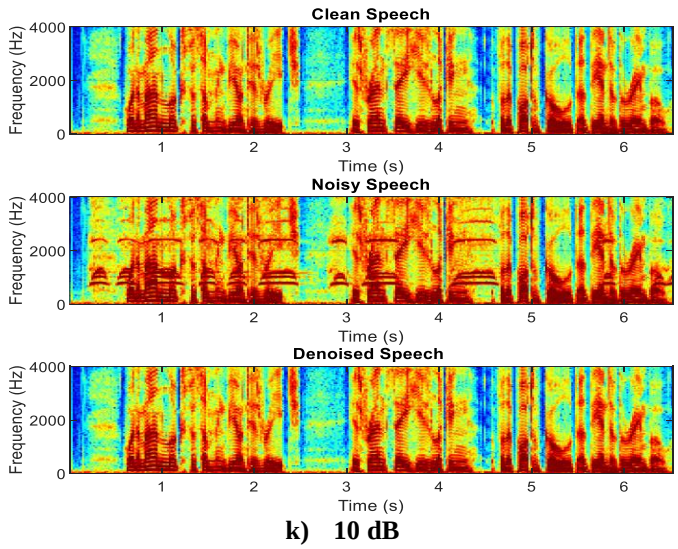
h) -5 dB



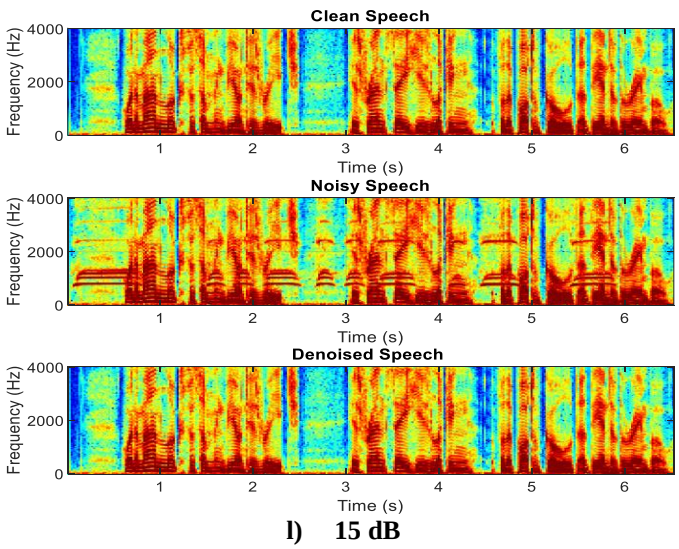
i) 0 dB



j) 5 dB



k) 10 dB



l) 15 dB

Figure 7 Modified FCRN - Spectrogram Images of Train Whistle Noise for various Noise Levels

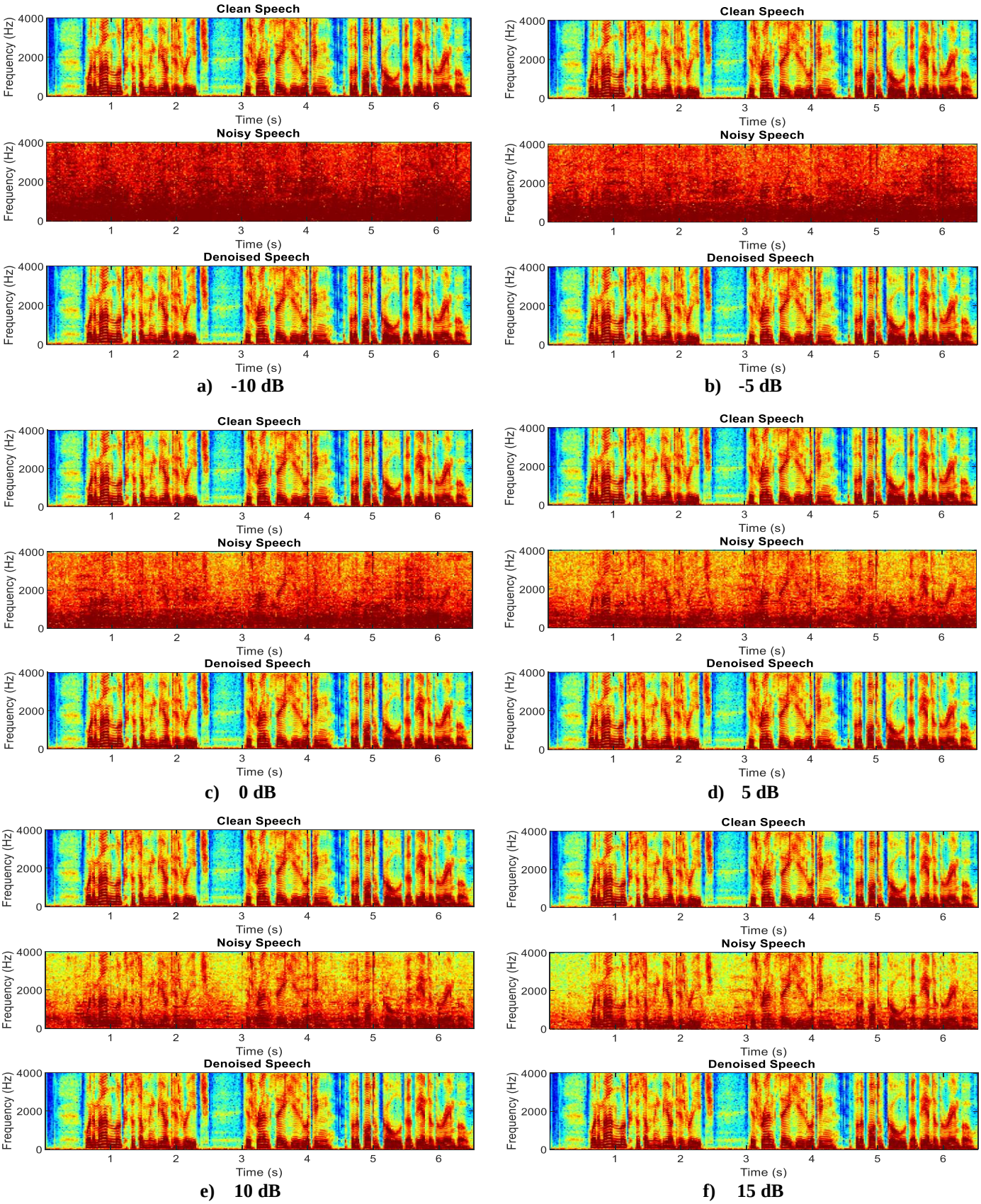


Figure 8 Modified FCRN - Spectrogram Images of Restaurant Noise for various Noise Levels

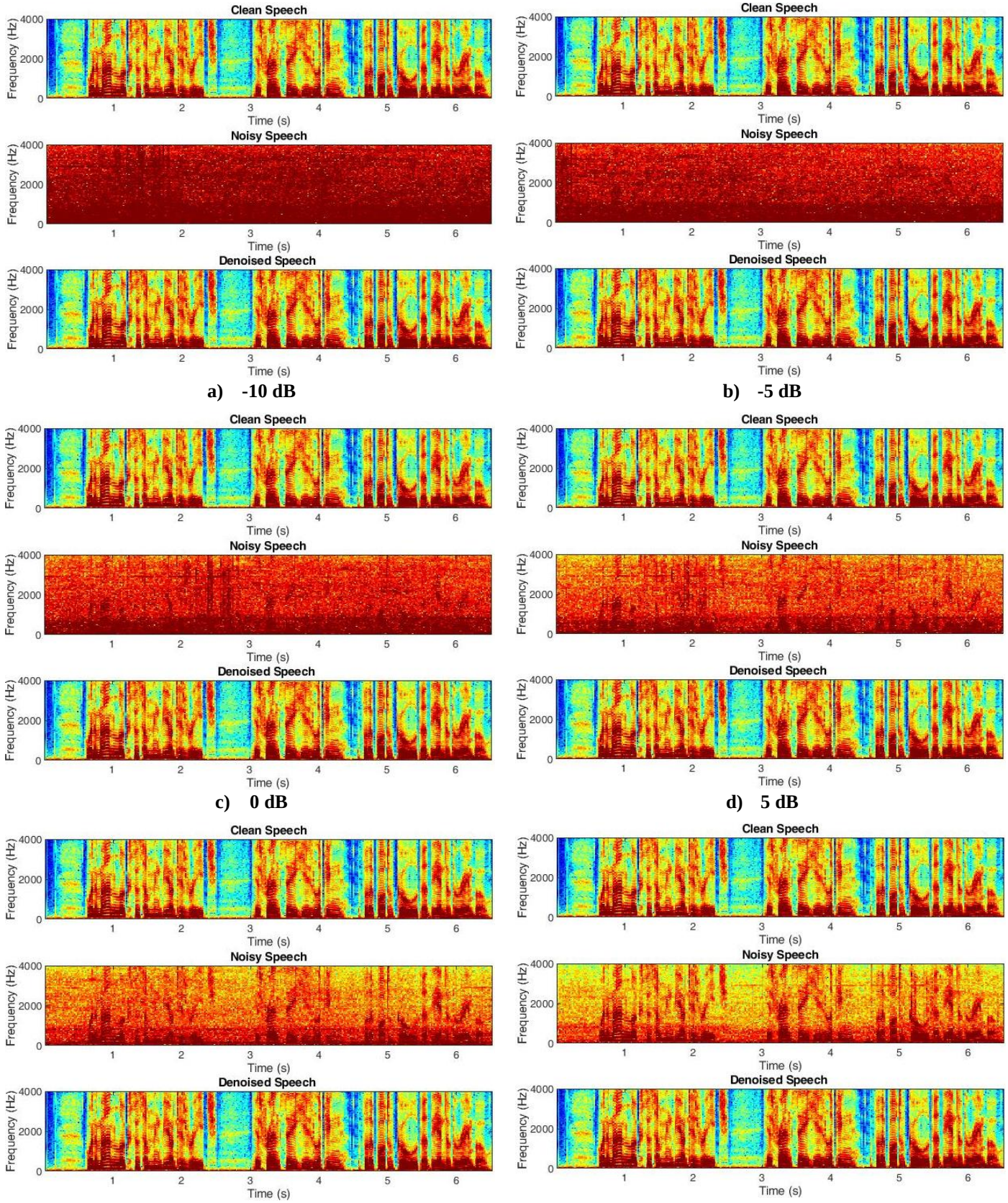
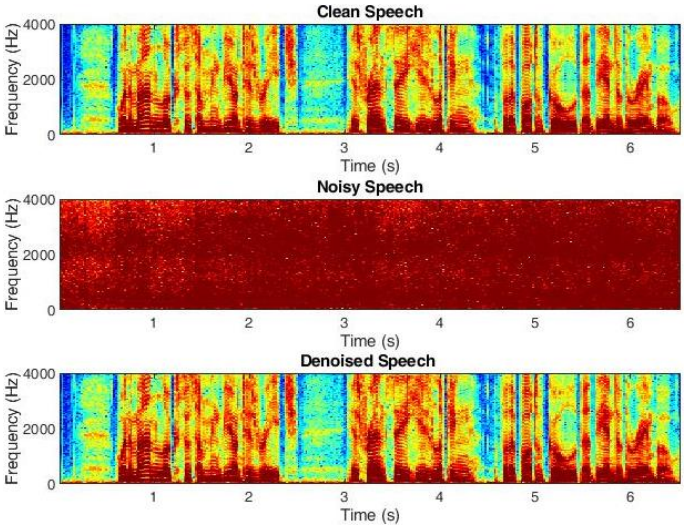
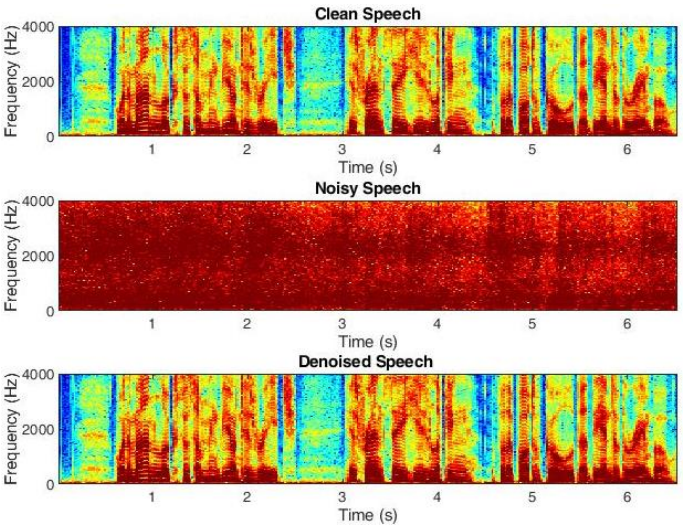


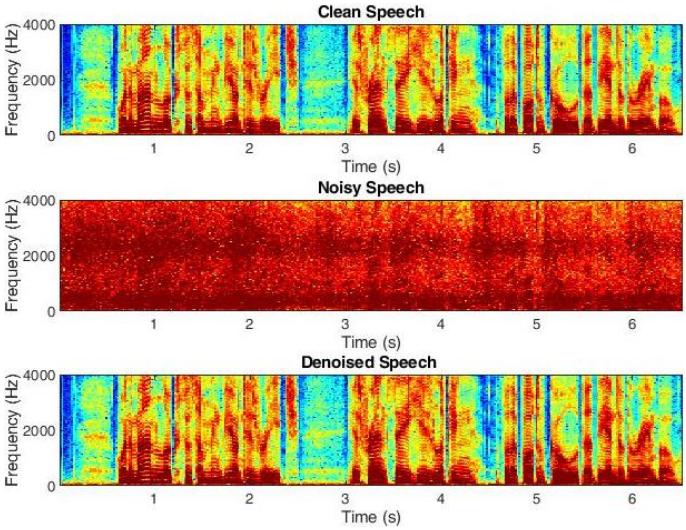
Figure 9 Modified FCRN - Spectrogram Images of Car Noise for various Noise Levels



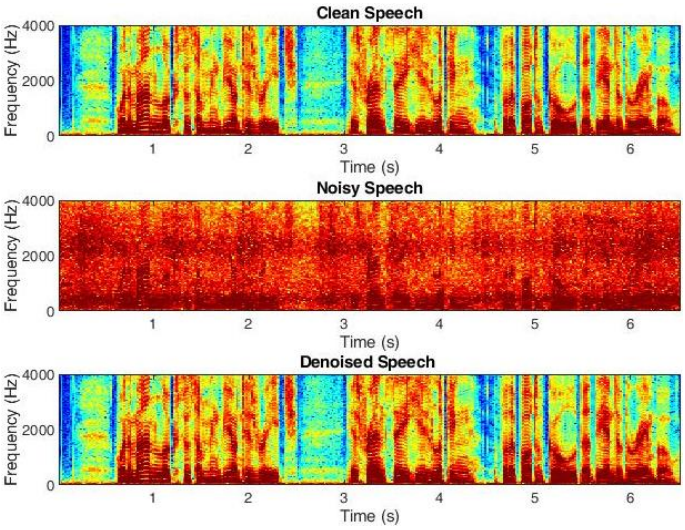
a) -10 dB



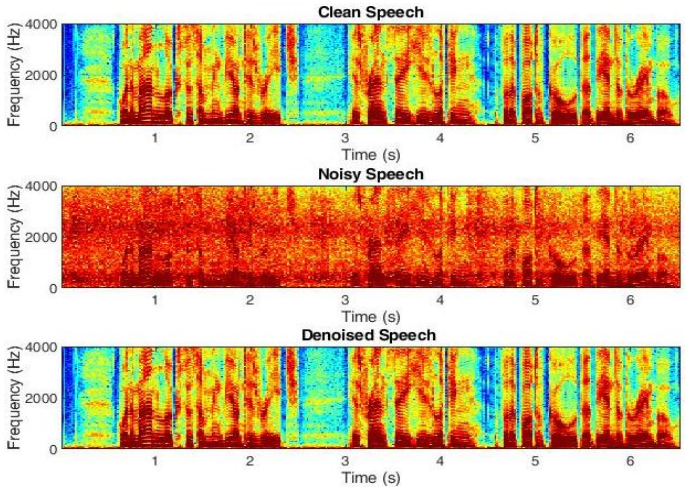
b) -5 dB



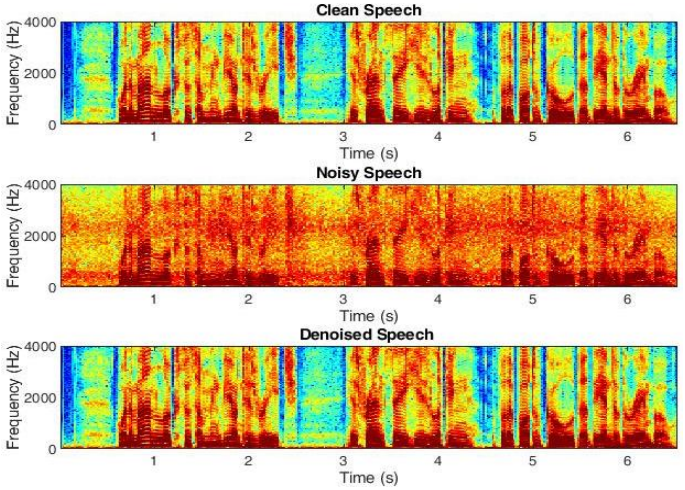
c) 0 dB



d) 5 dB



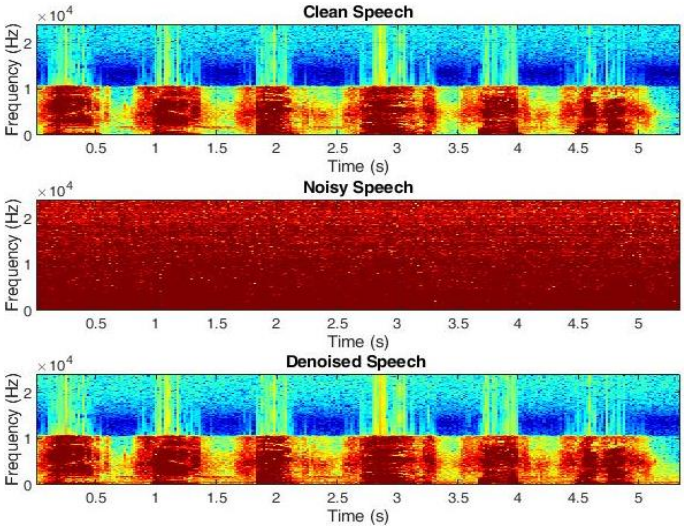
e) 10 dB



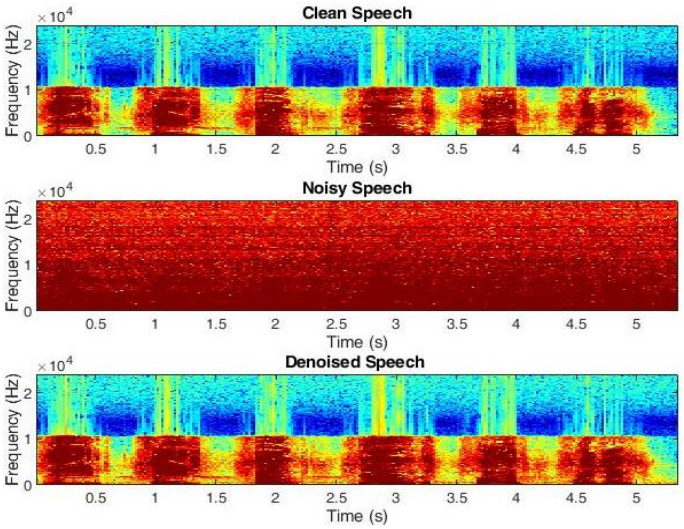
f) 15 dB

Figure 10 Modified FCRN - Spectrogram Images of Subway Noise for various Noise Levels

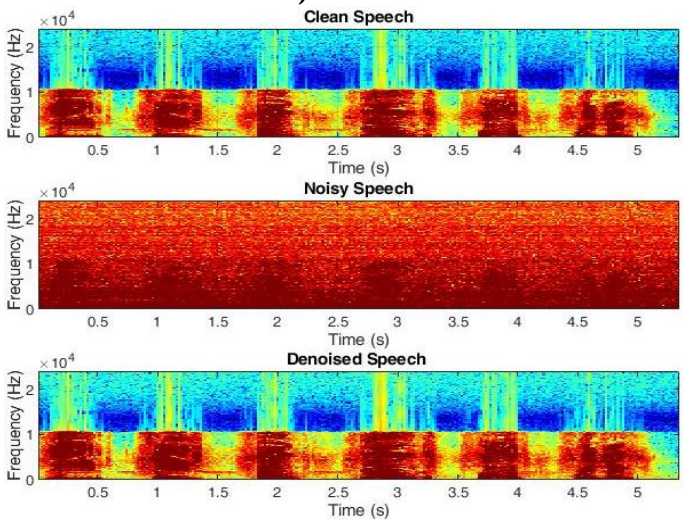
APPENDIX 5 (Spectrograms of Alaryngeal Speech)



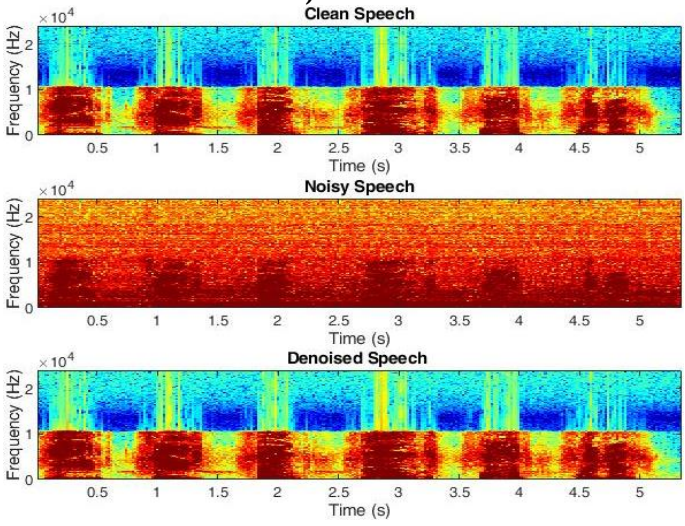
a) -10 dB



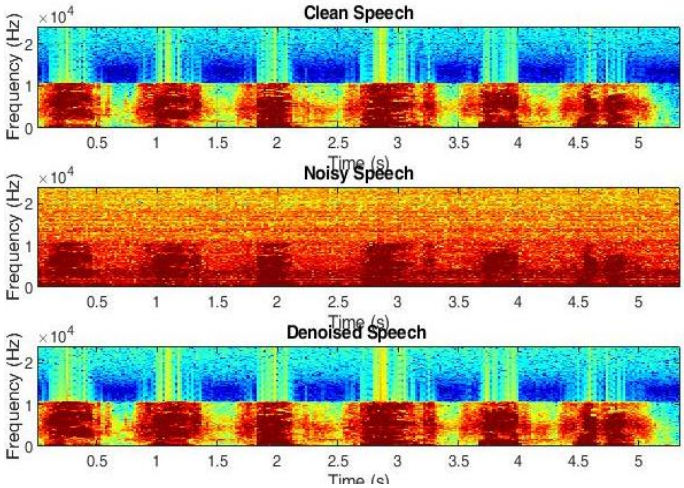
b) -5 dB



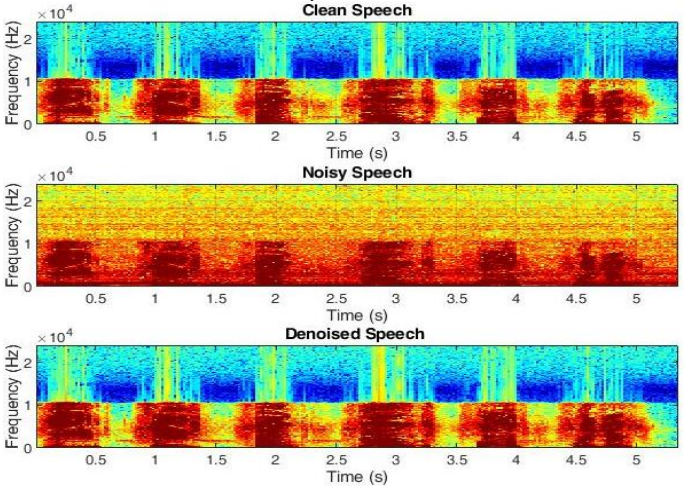
c) 0 dB



d) 5 dB

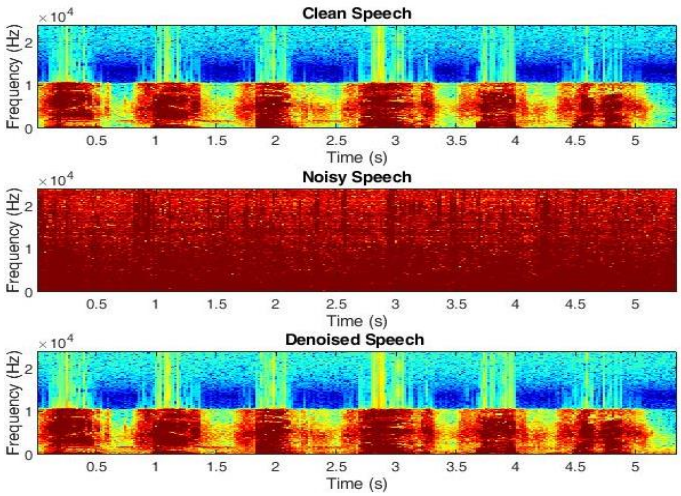


e) 10 dB

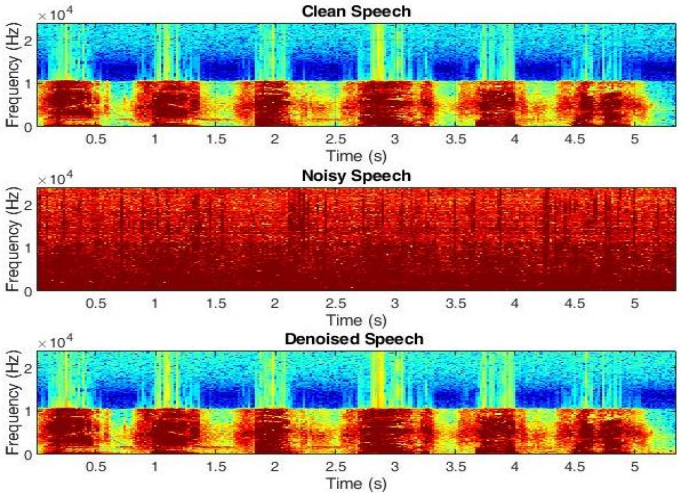


f) 15 dB

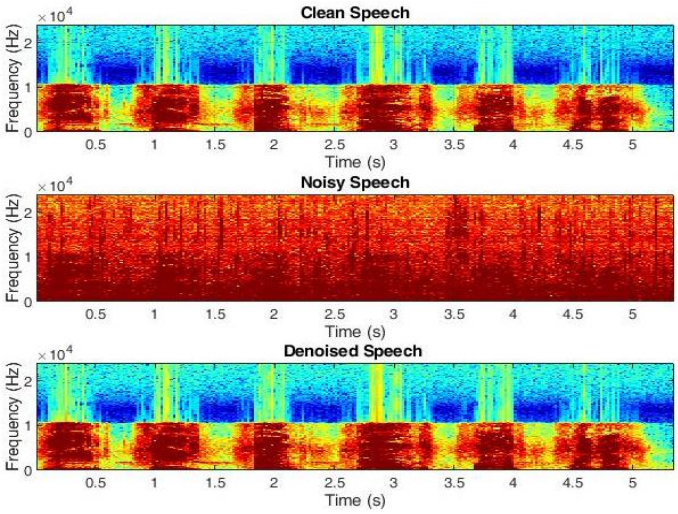
Figure 1 DFNN – Spectrogram Images of Washing Machine Noise for Alaryngeal Speech at various Noise Levels



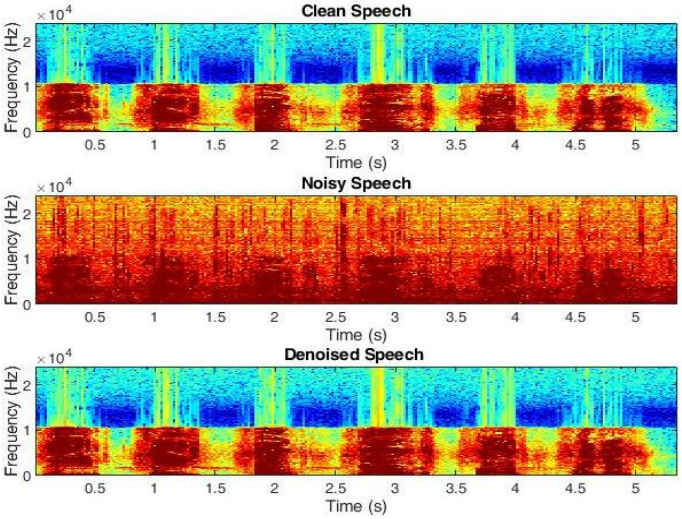
a) -10 dB



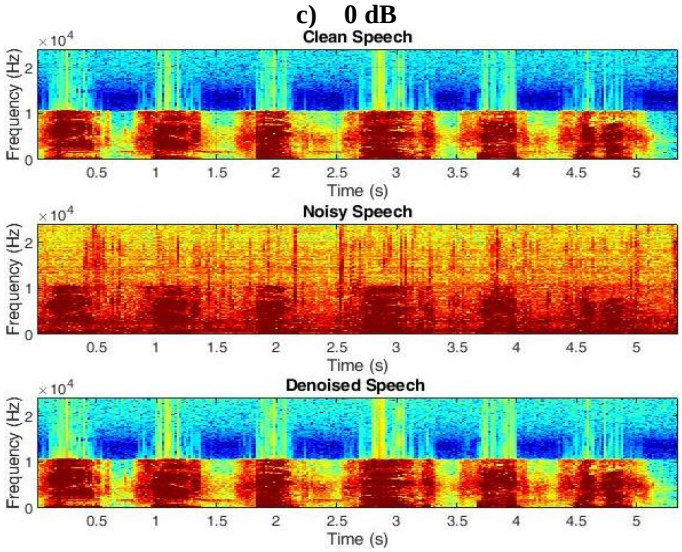
b) -5 dB



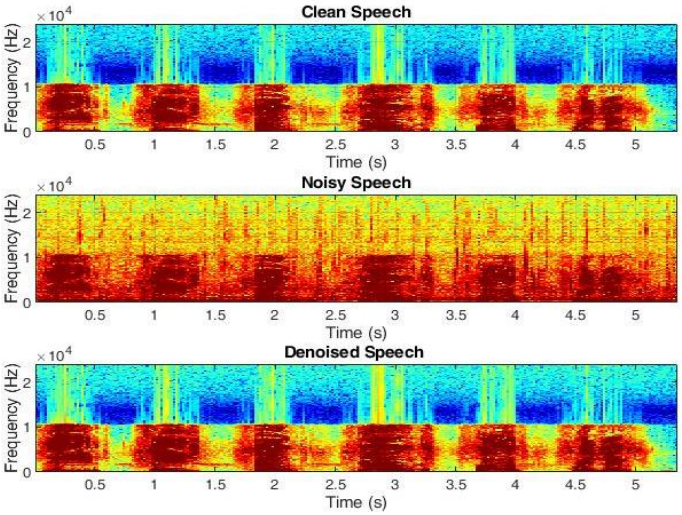
c) 0 dB



d) 5 dB



e) 10 dB



f) 15 dB

Figure 2 DFNN – Spectrogram Images of Rainbow Noise for Alaryngeal Speech at various Noise Levels

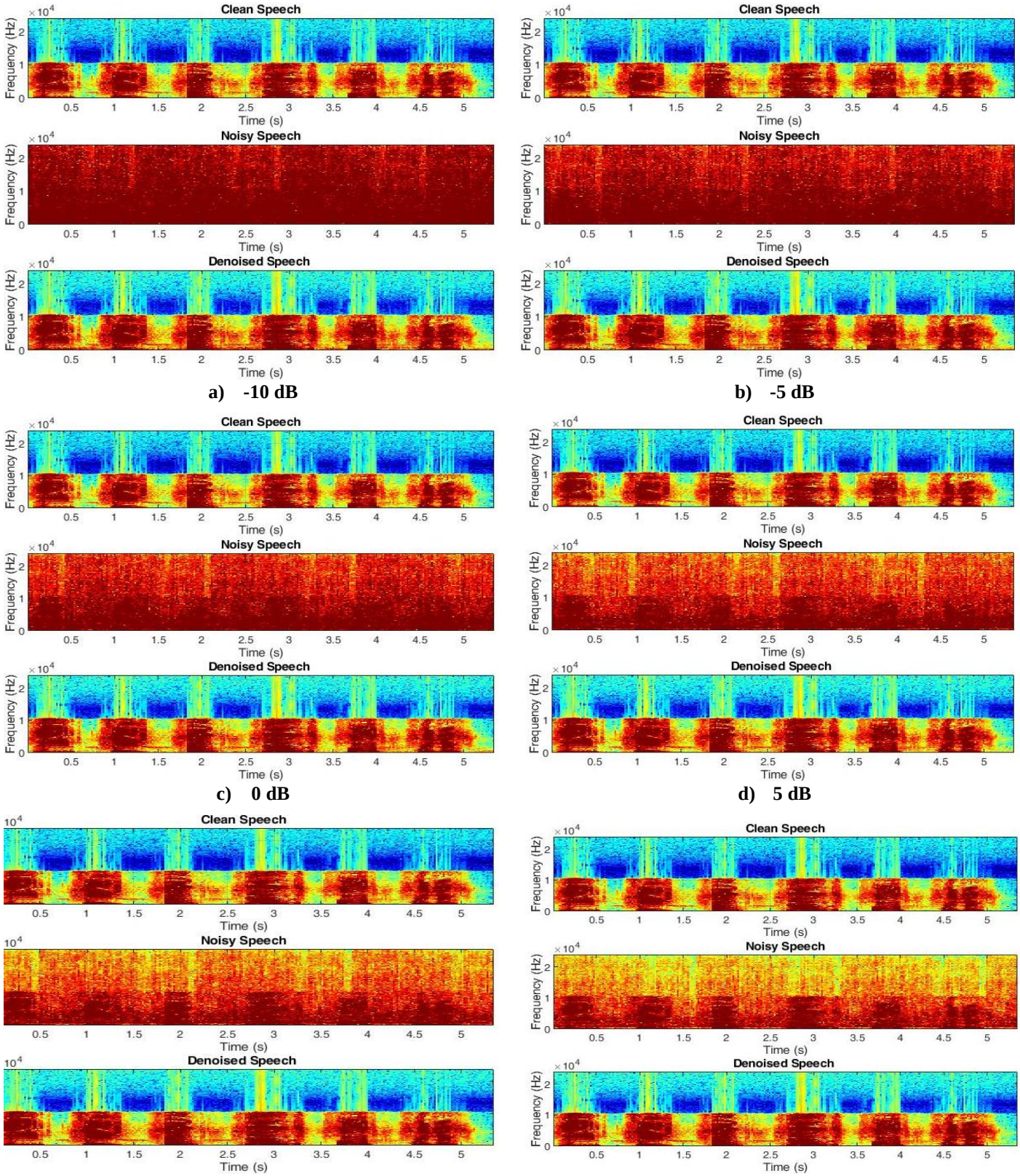
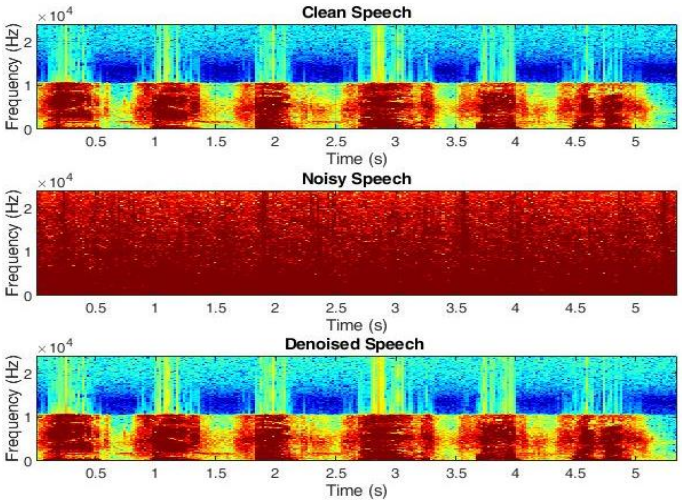
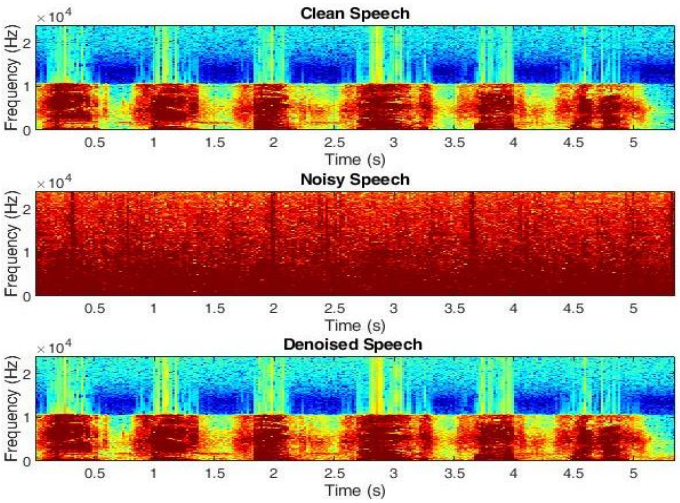


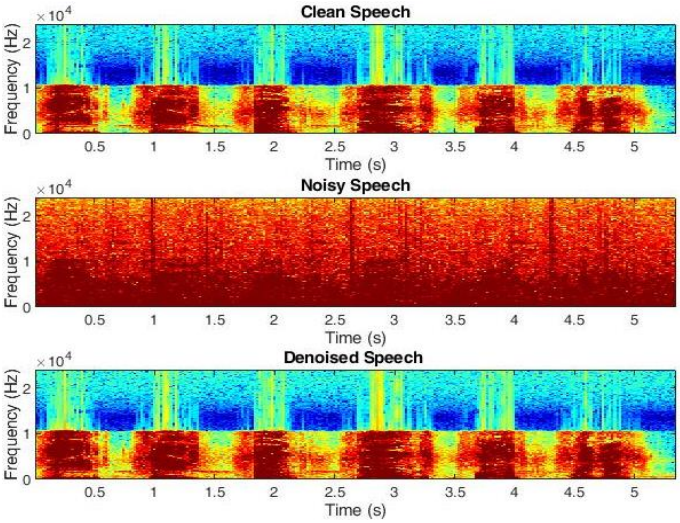
Figure 3 DFNN – Spectrogram Images of Babble Noise for Alaryngeal Speech at various Noise Levels



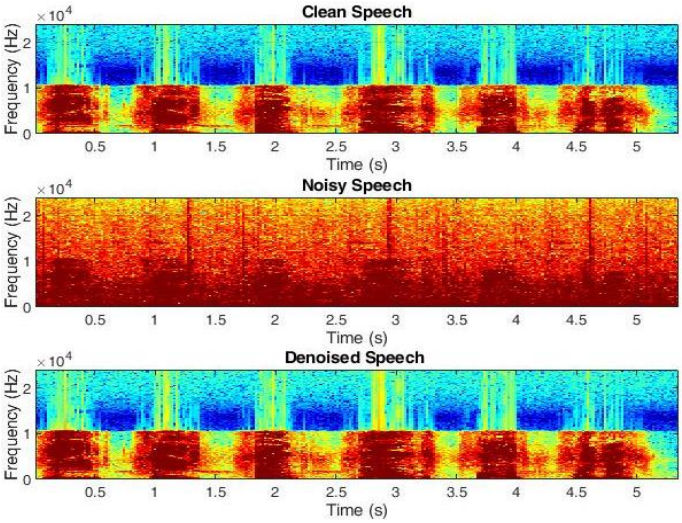
a) -10 dB



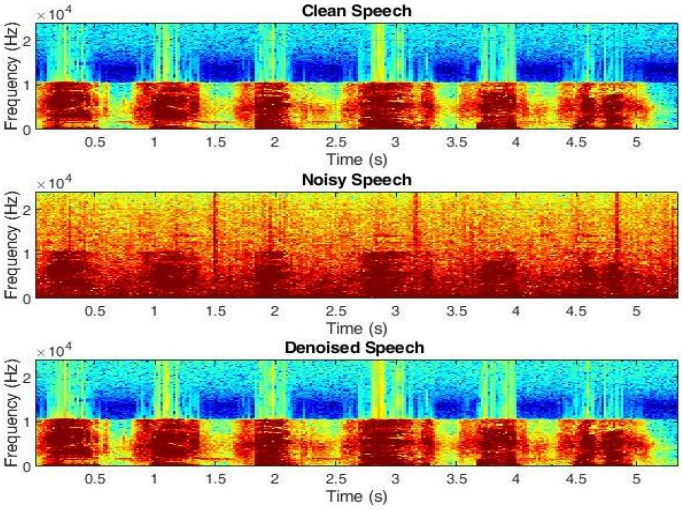
b) -5 dB



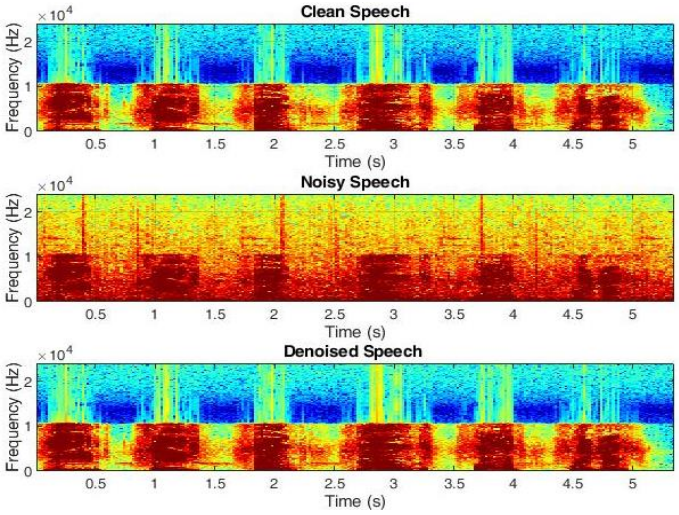
c) 0 dB



d) 5 dB

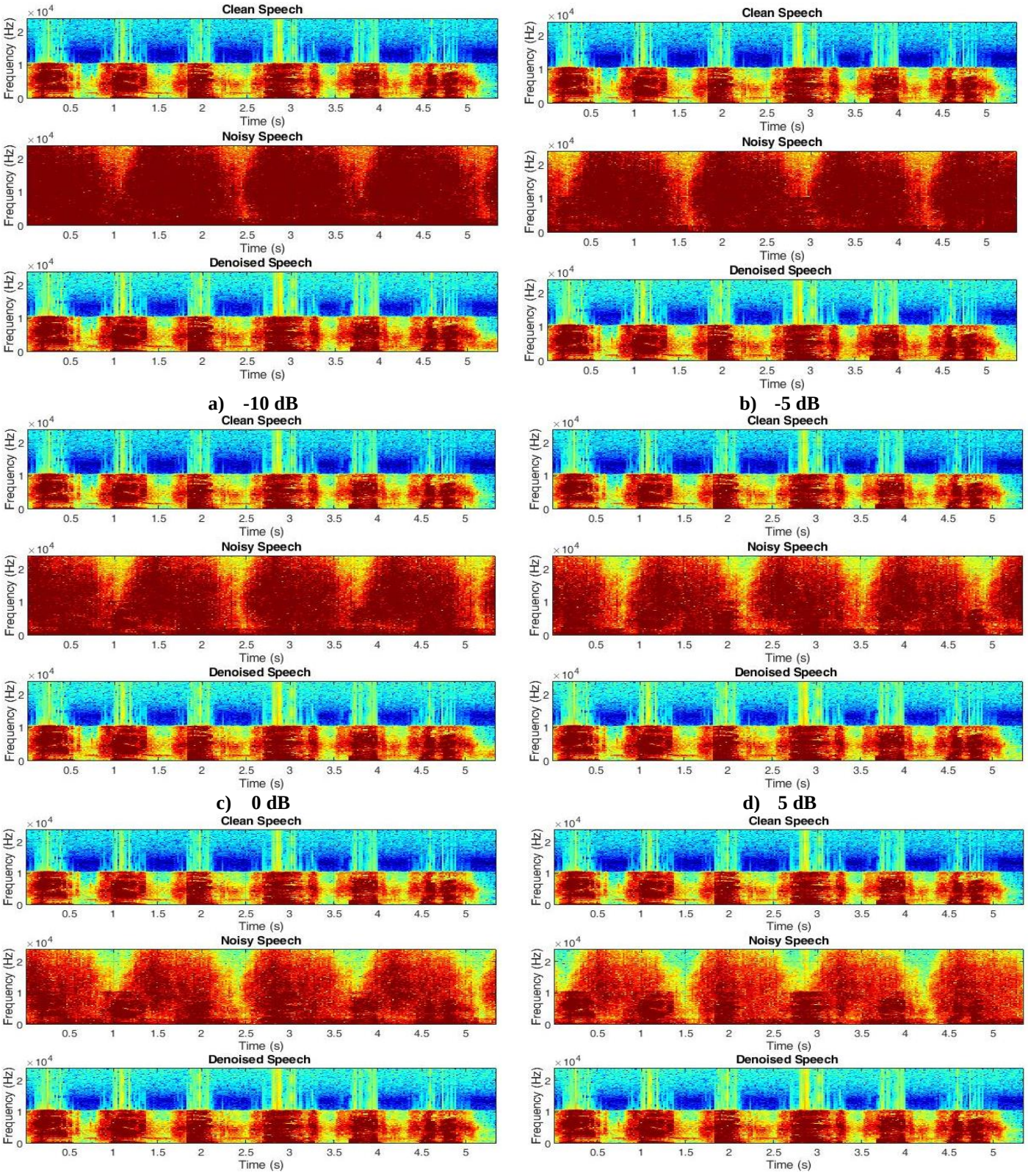


e) 10 dB



f) 15 dB

Figure 4 DFNN – Spectrogram Images of Airport Noise for Alaryngeal Speech at various Noise Levels



a) -10 dB **b) -5 dB**

c) 0 dB **d) 5 dB**

e) 10 dB **f) 15 dB**

Figure 5 DFNN – Spectrogram Images of Jetplane Noise for Alaryngeal Speech at various Noise Levels

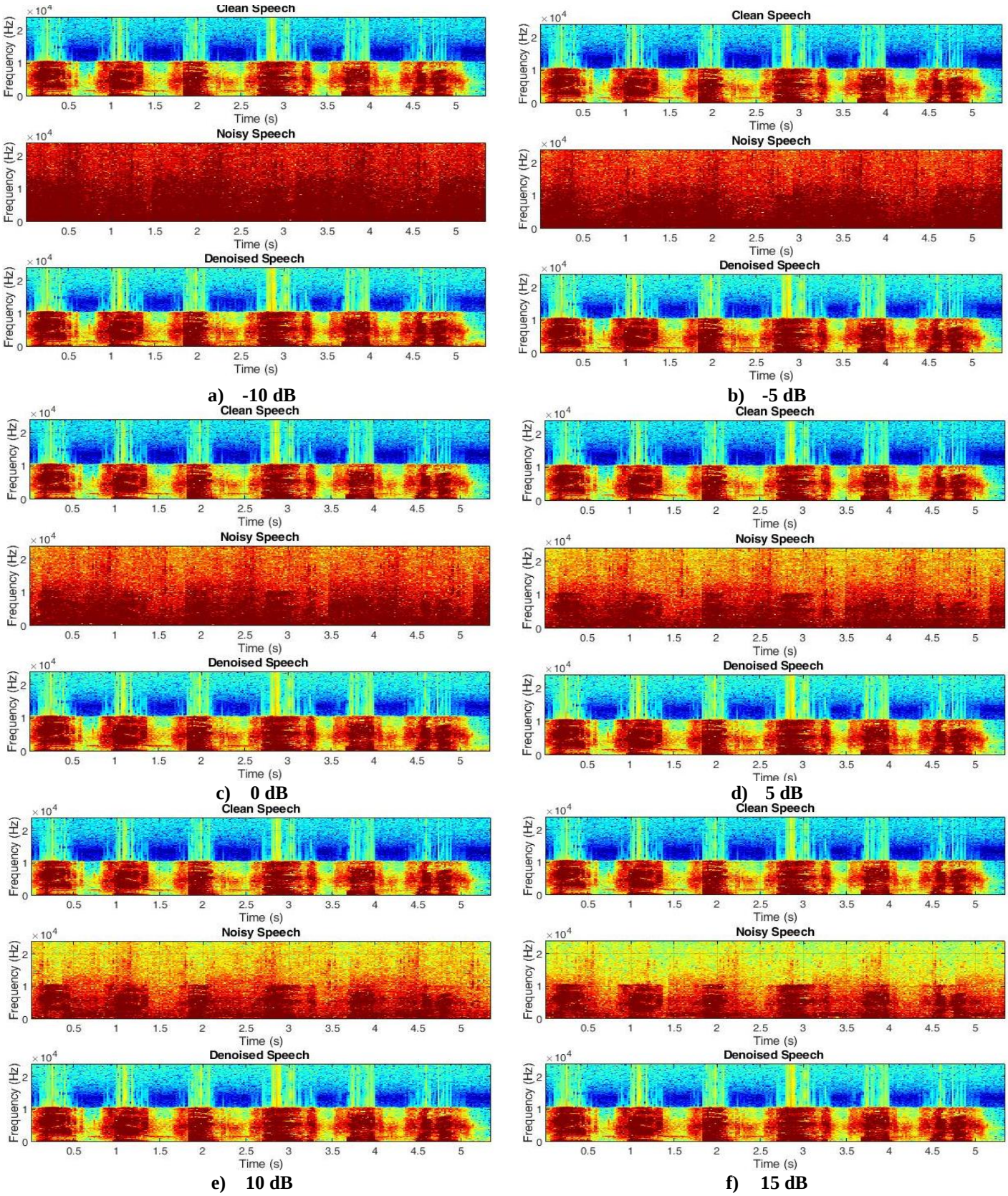
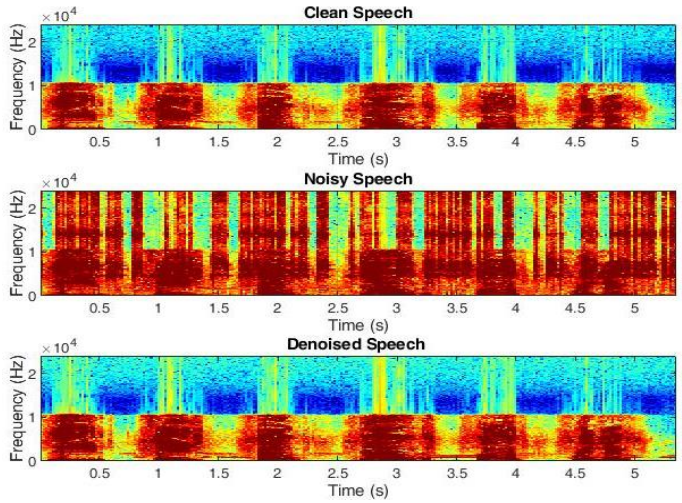
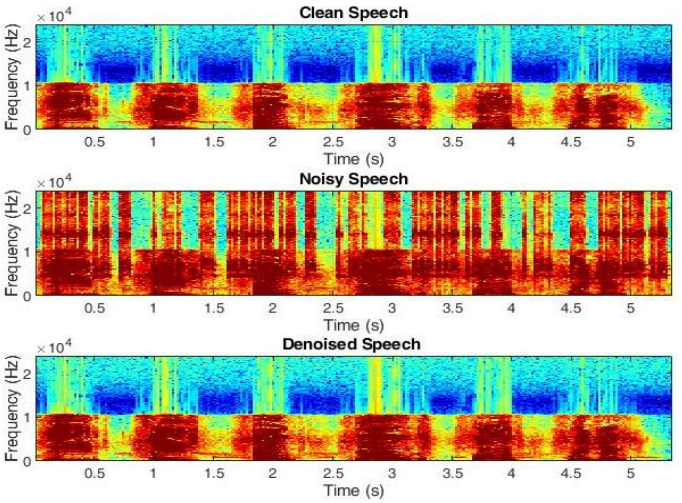


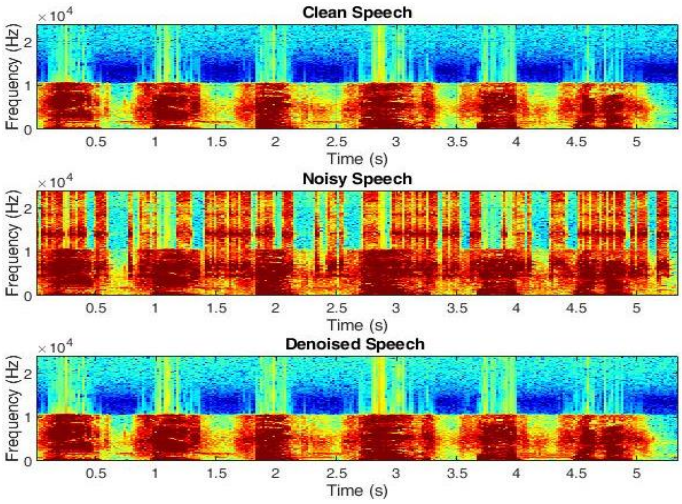
Figure 6 DFNN – Spectrogram Images of Street Noise for Alaryngeal Speech at various Noise Levels



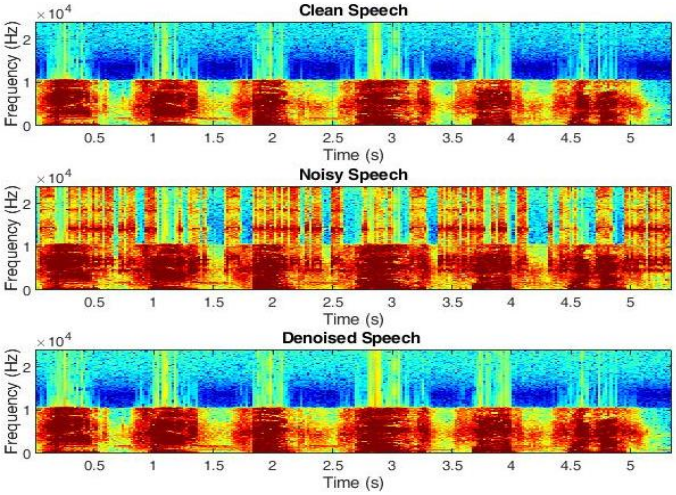
a) -10 dB



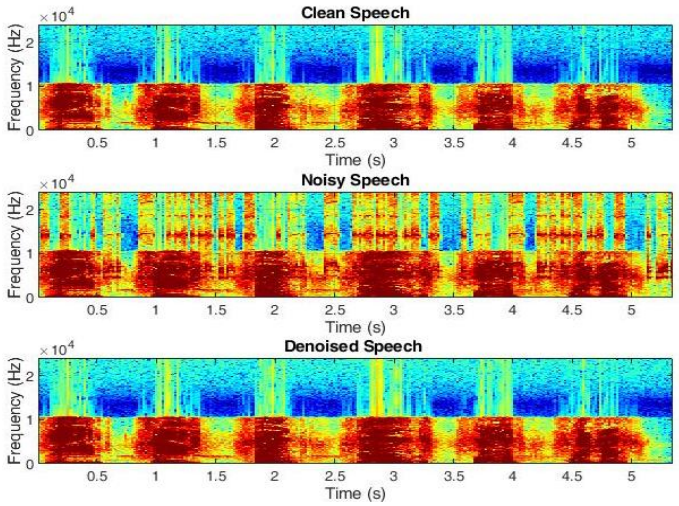
b) -5 dB



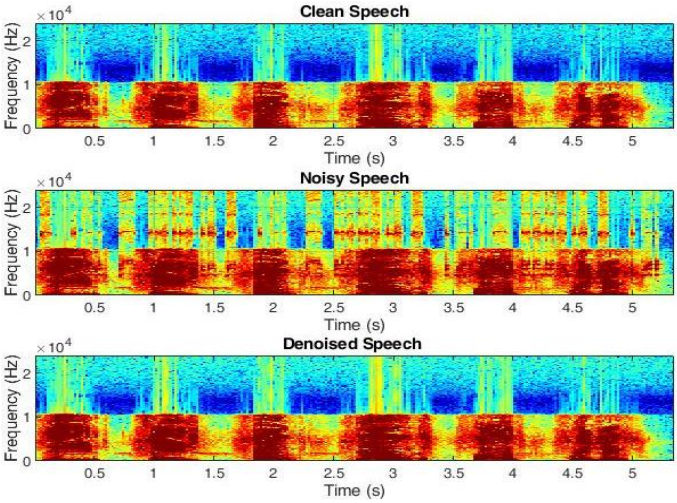
c) 0 dB



d) 5 dB

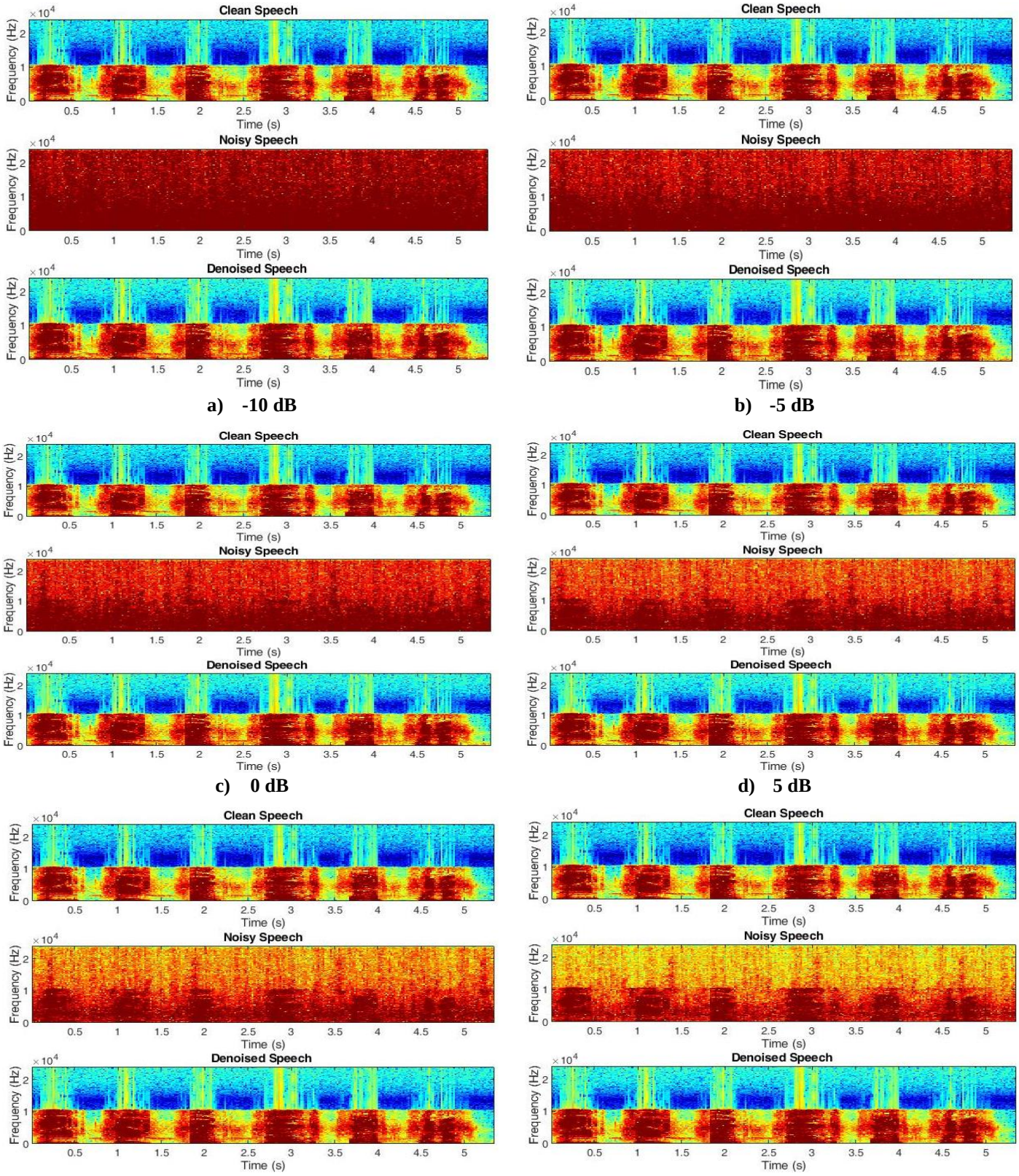


e) 10 dB



f) 15 dB

Figure 7 DFNN – Spectrogram Images of Train Whistle Noise for Alaryngeal Speech at various Noise Levels



a) -10 dB b) -5 dB

c) 0 dB d) 5 dB

e) 10 dB f) 15 dB

Figure 8 DFNN – Spectrogram Images of Restaurant Noise for Alaryngeal Speech at various Noise Levels

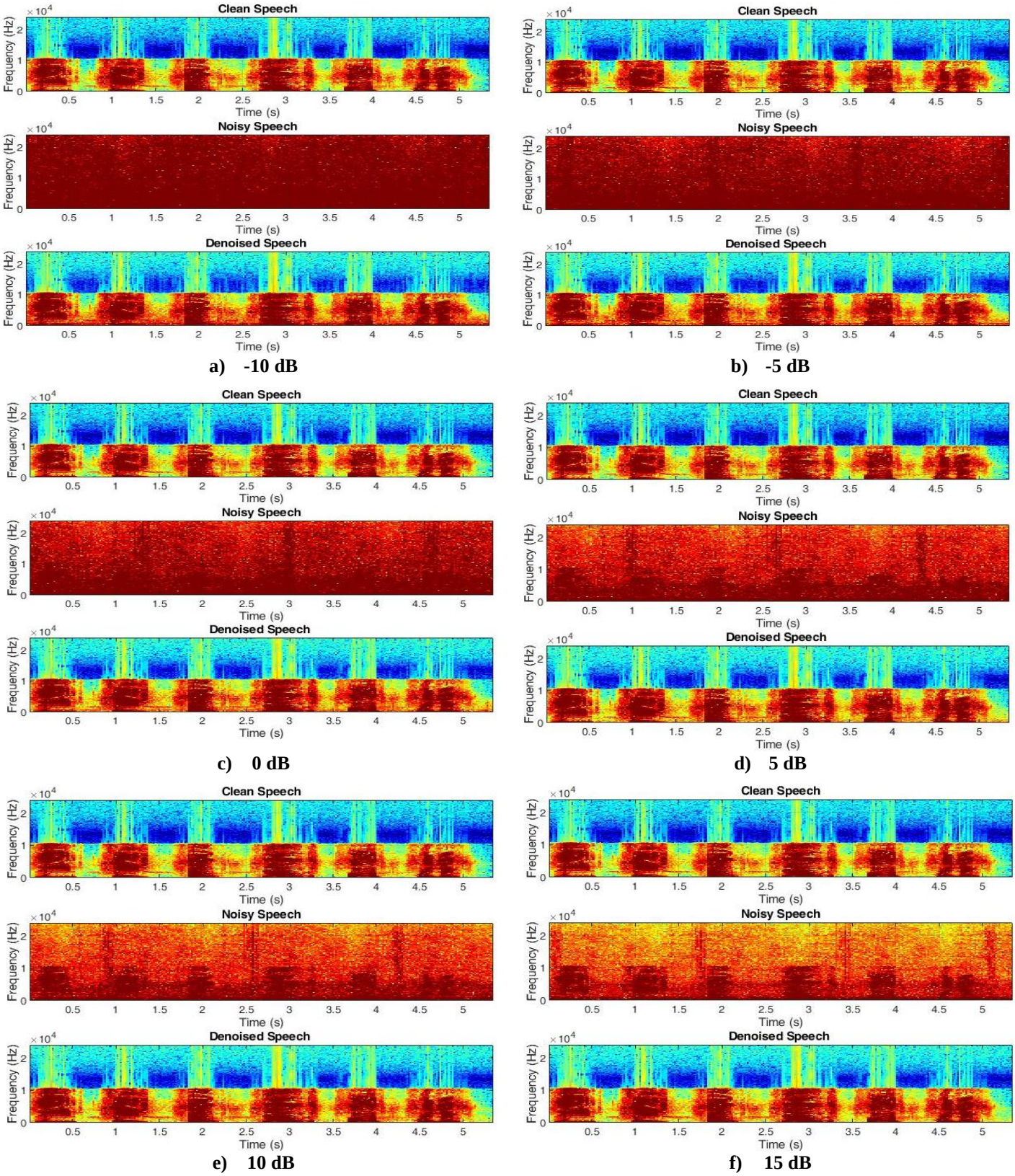
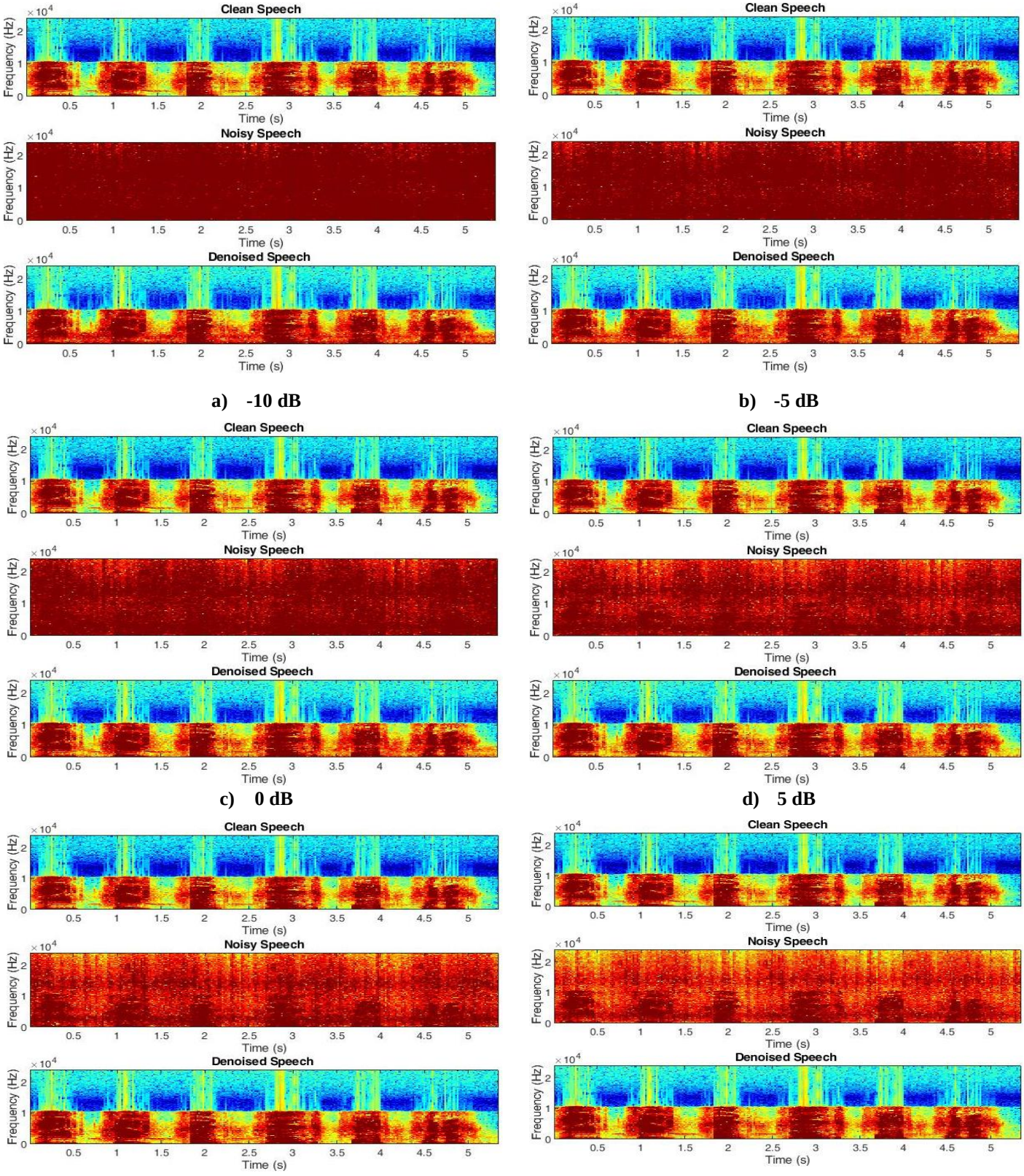


Figure 9 DFNN – Spectrogram Images of Car Noise for Alaryngeal Speech at various Noise Levels

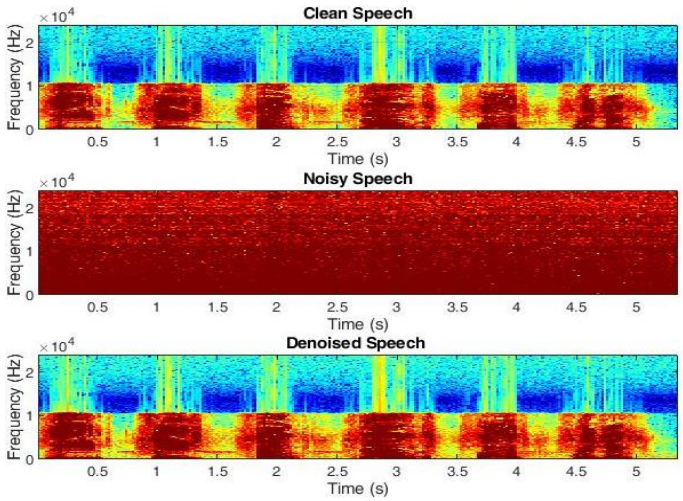


a) -10 dB b) -5 dB

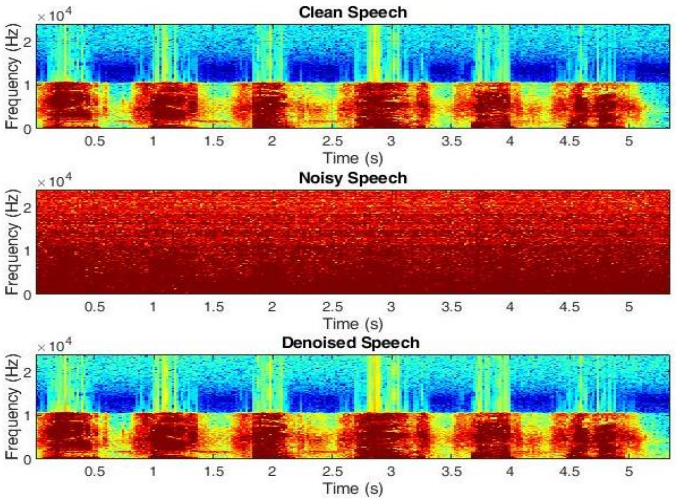
c) 0 dB d) 5 dB

e) 10 dB f) 15 dB

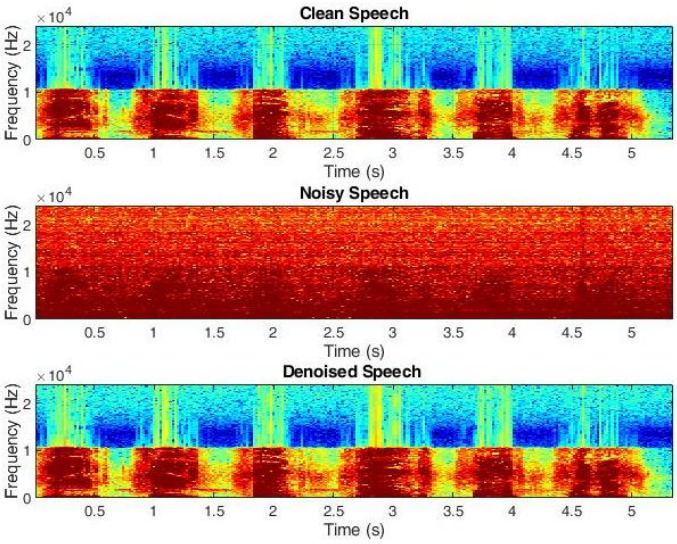
Figure 10 DFNN – Spectrogram Images of Subway Noise for Alaryngeal Speech at various Noise Levels



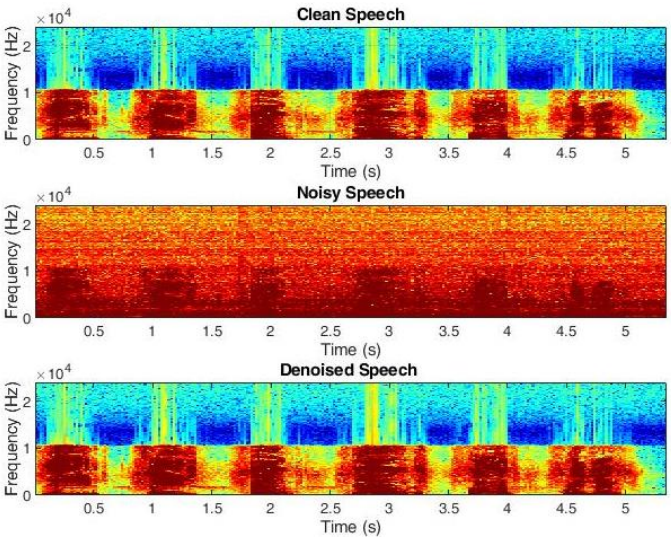
a) -10 dB



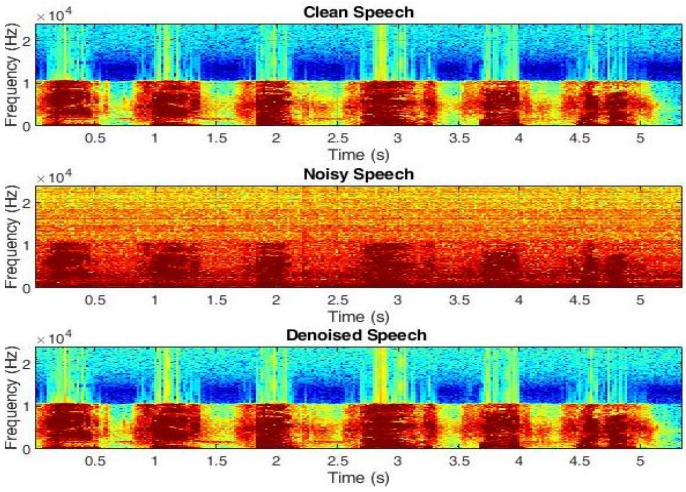
b) -5 dB



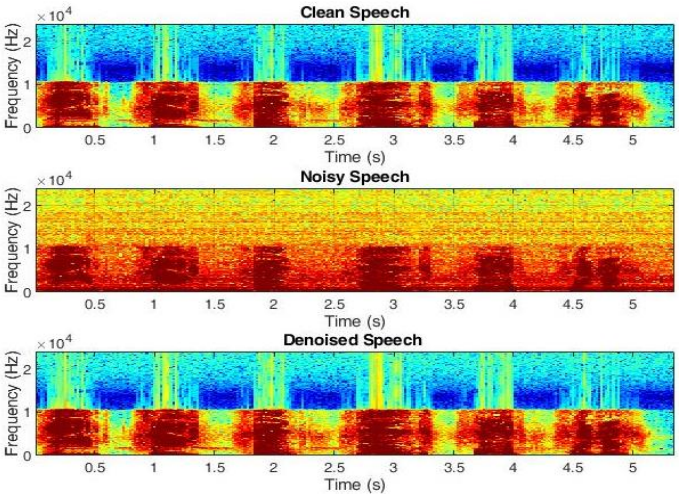
c) 0 dB



d) 5 dB

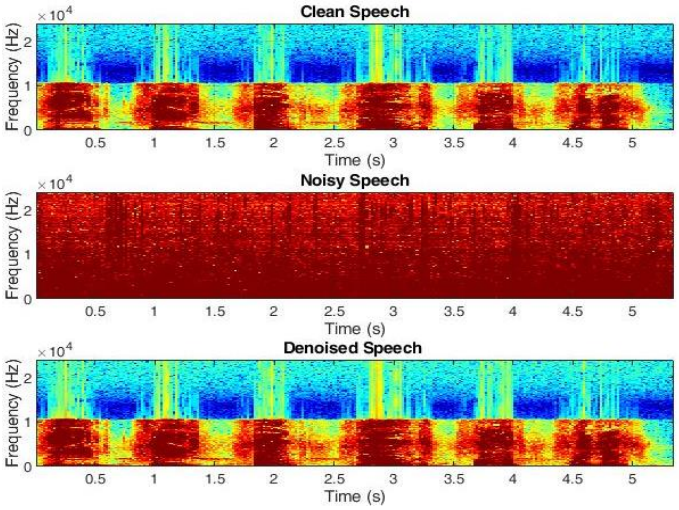


e) 10 dB

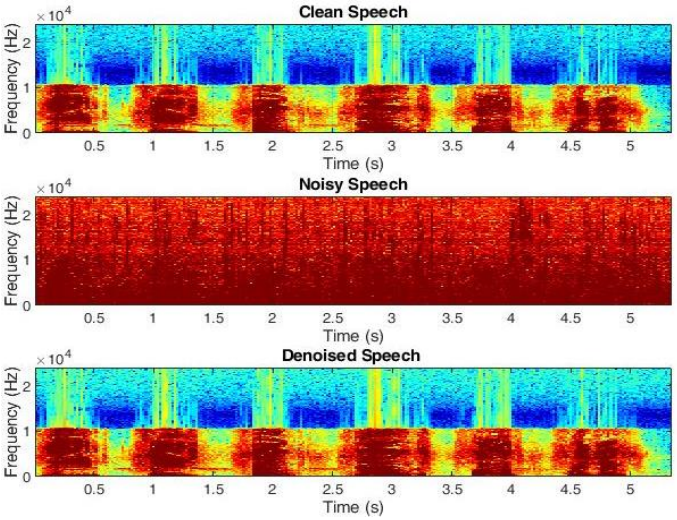


f) 15 dB

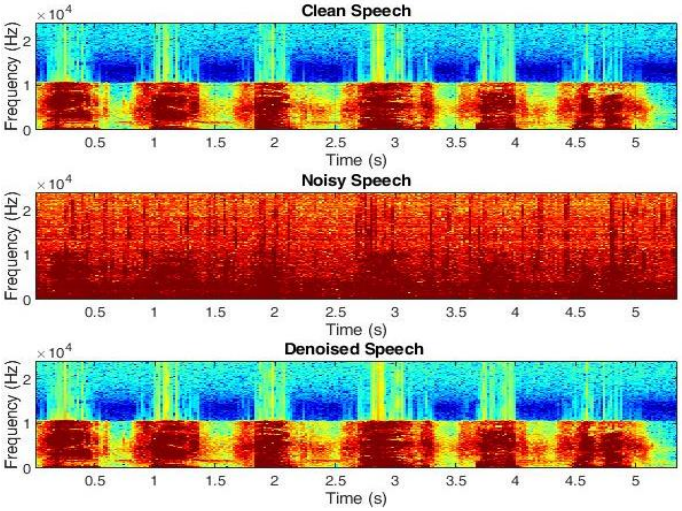
Figure 11 Deep CNN – Spectrogram Images of Washing Machine Noise for Alaryngeal Speech at various Noise Levels



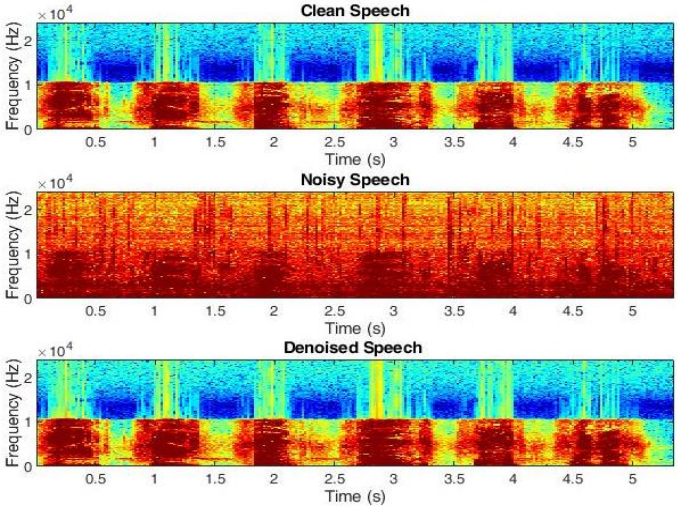
a) -10 dB



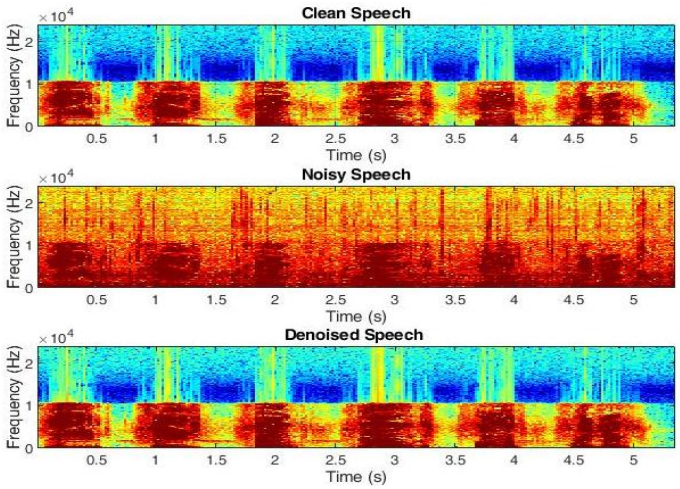
b) -5 dB



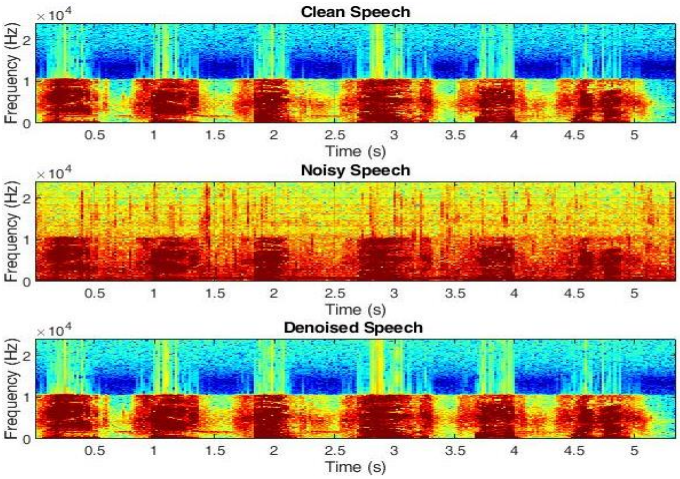
c) 0 dB



d) 5 dB

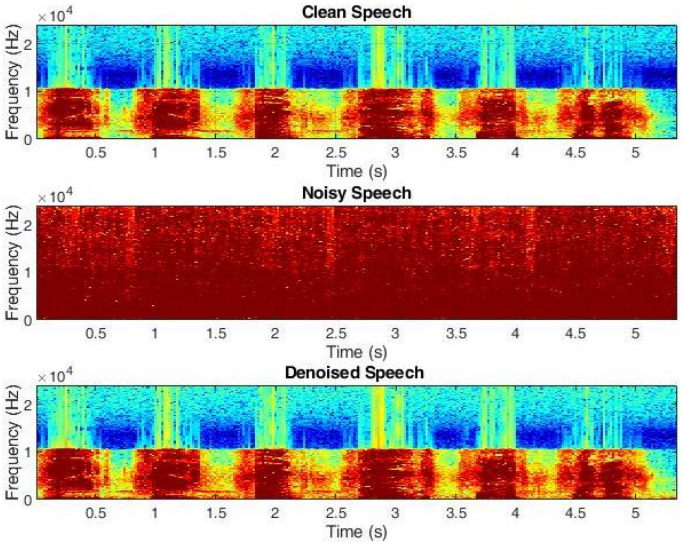


e) 10 dB

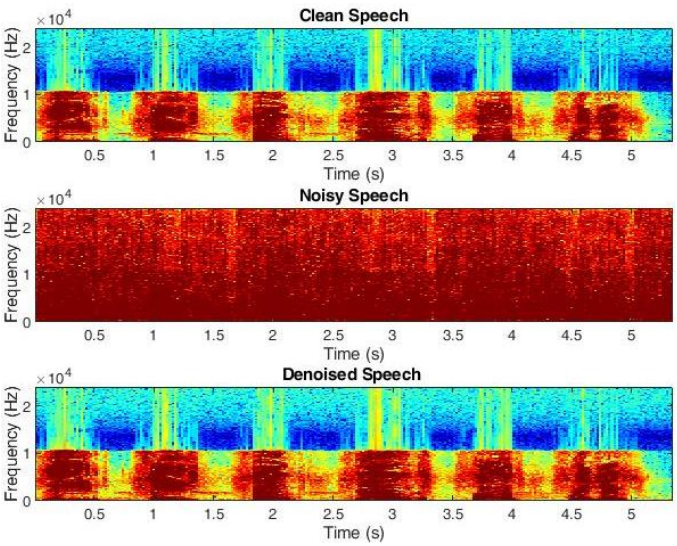


f) 15 dB

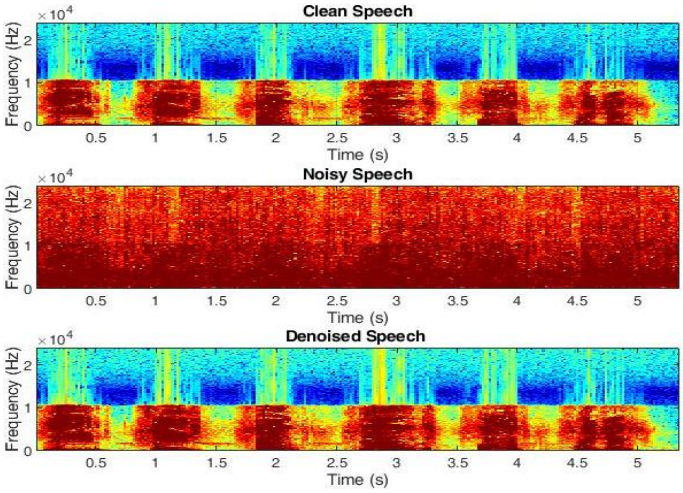
Figure 12 Deep CNN – Spectrogram Images of Rainbow Noise for Alaryngeal Speech at various Noise Levels



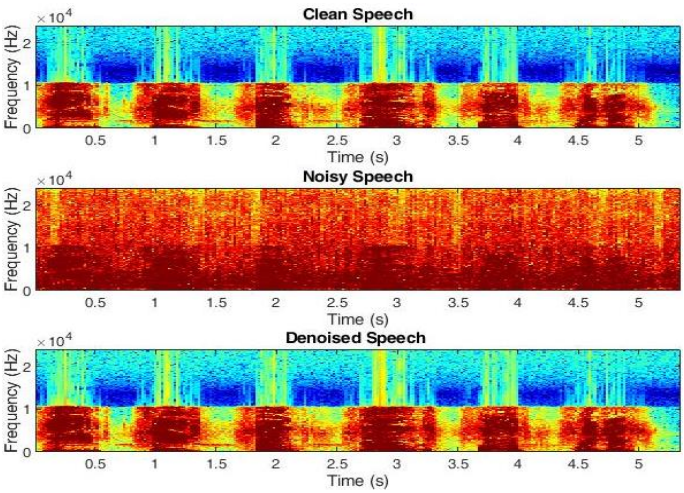
a) -10 dB



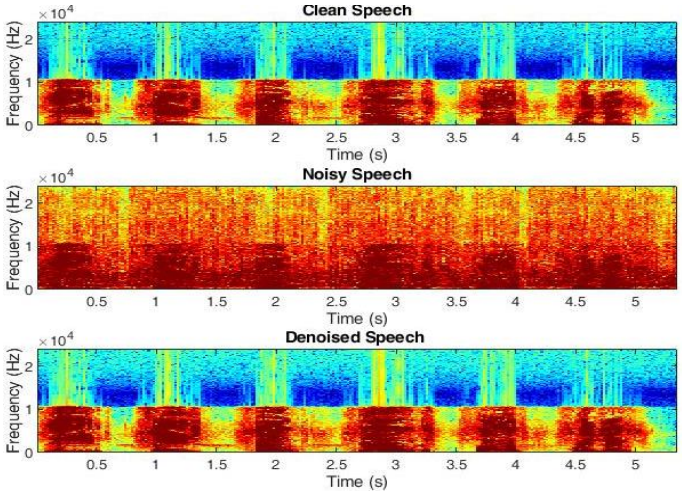
b) -5 dB



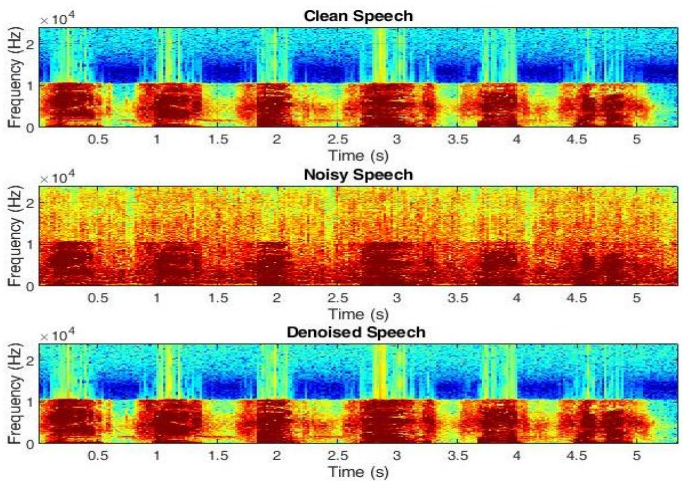
c) 0 dB



d) 5 dB

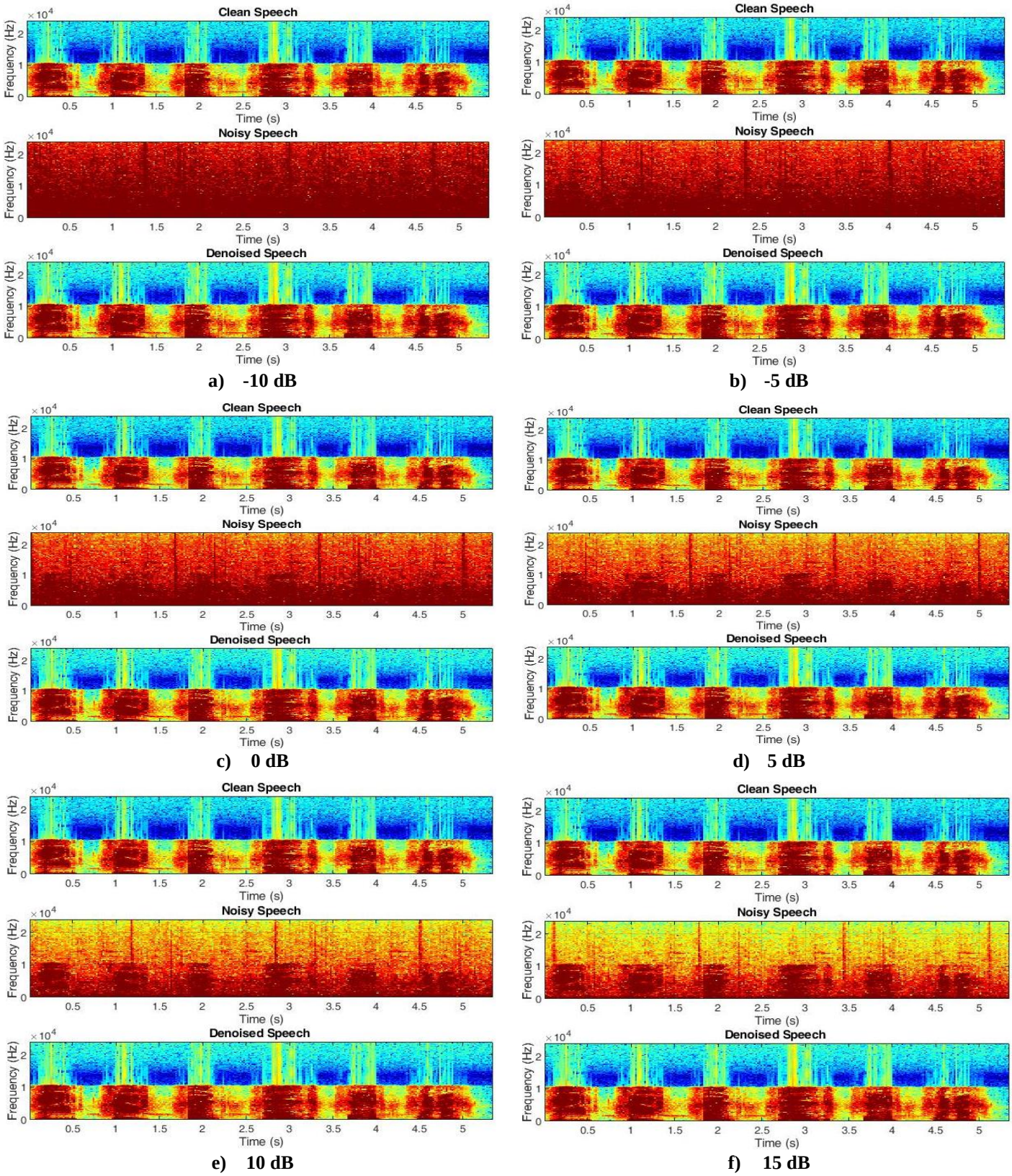


e) 10 dB



f) 15 dB

Figure 13 Deep CNN – Spectrogram Images of Babble Noise for Alaryngeal Speech at various Noise Levels

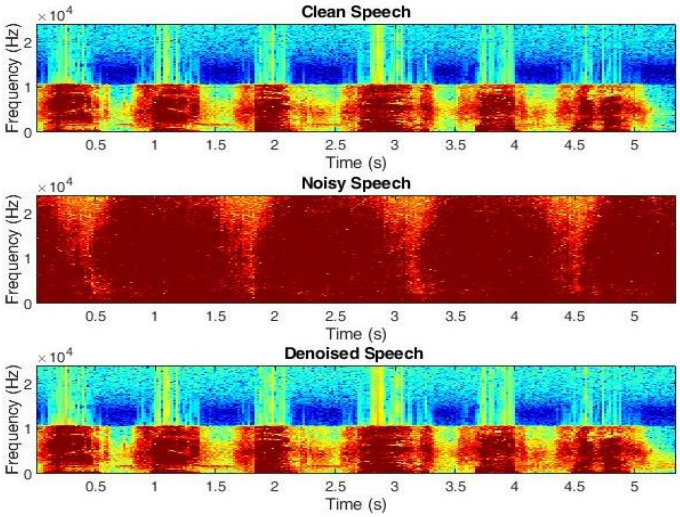


a) -10 dB **b) -5 dB**

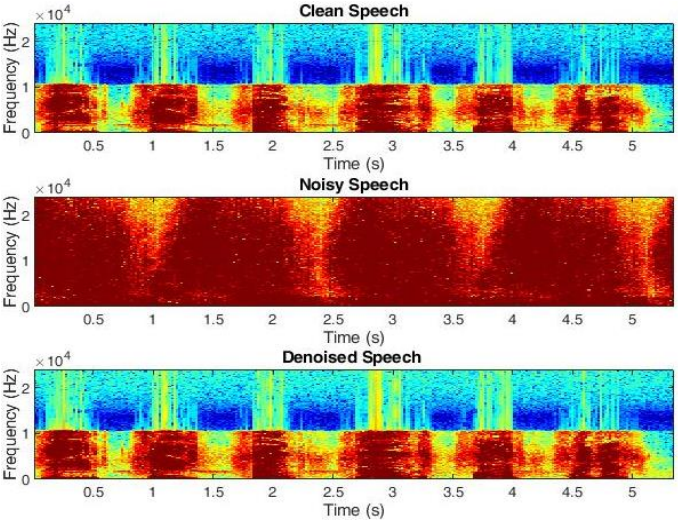
c) 0 dB **d) 5 dB**

e) 10 dB **f) 15 dB**

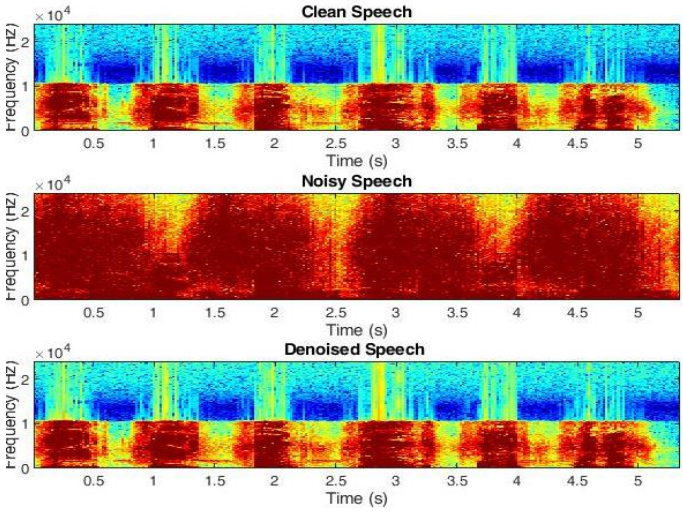
Figure 14 Deep CNN – Spectrogram Images of Airport Noise for Alaryngeal Speech at various Noise Levels



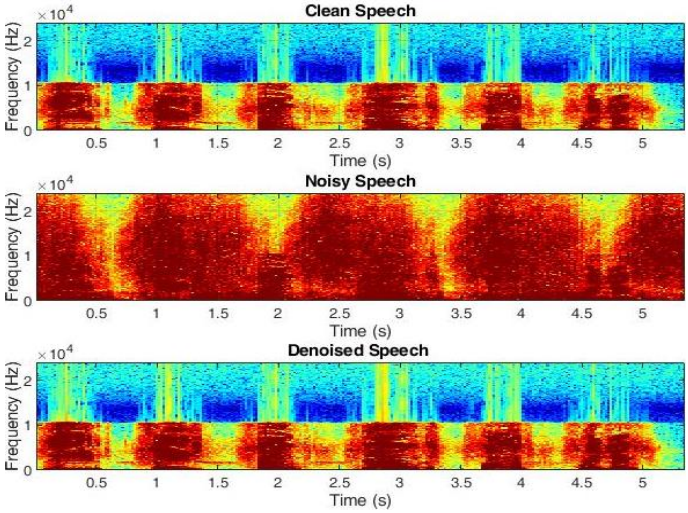
a) -10 dB



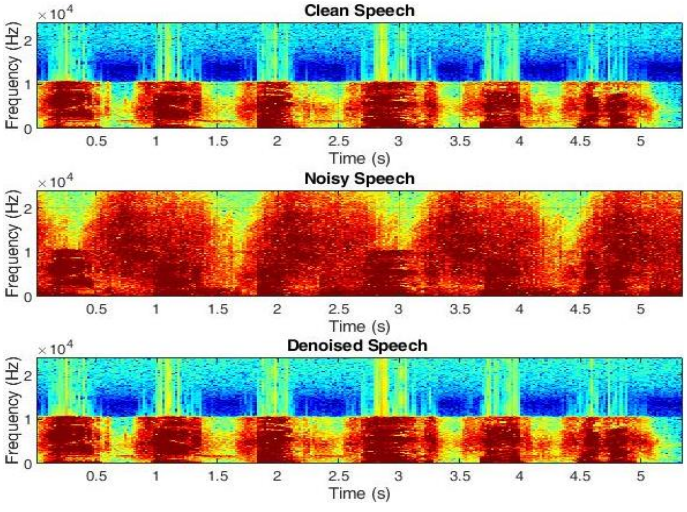
b) -5 dB



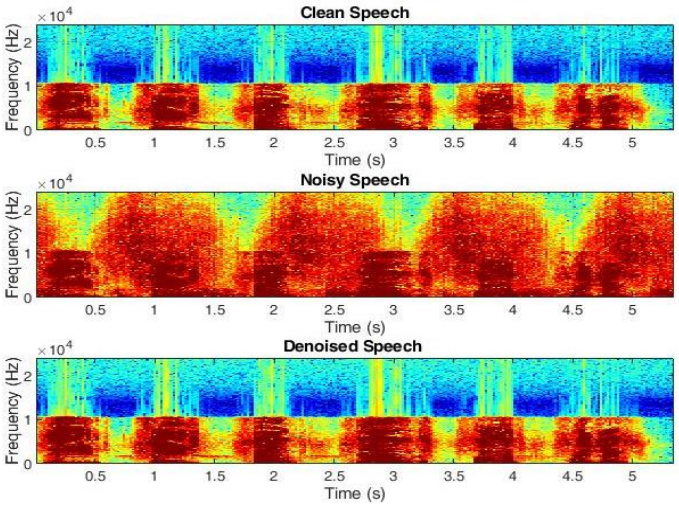
c) 0 dB



d) 5 dB



e) 10 dB



f) 15 dB

Figure 15 Deep CNN – Spectrogram Images of Jet plane Noise for Alaryngeal Speech at various Noise Levels

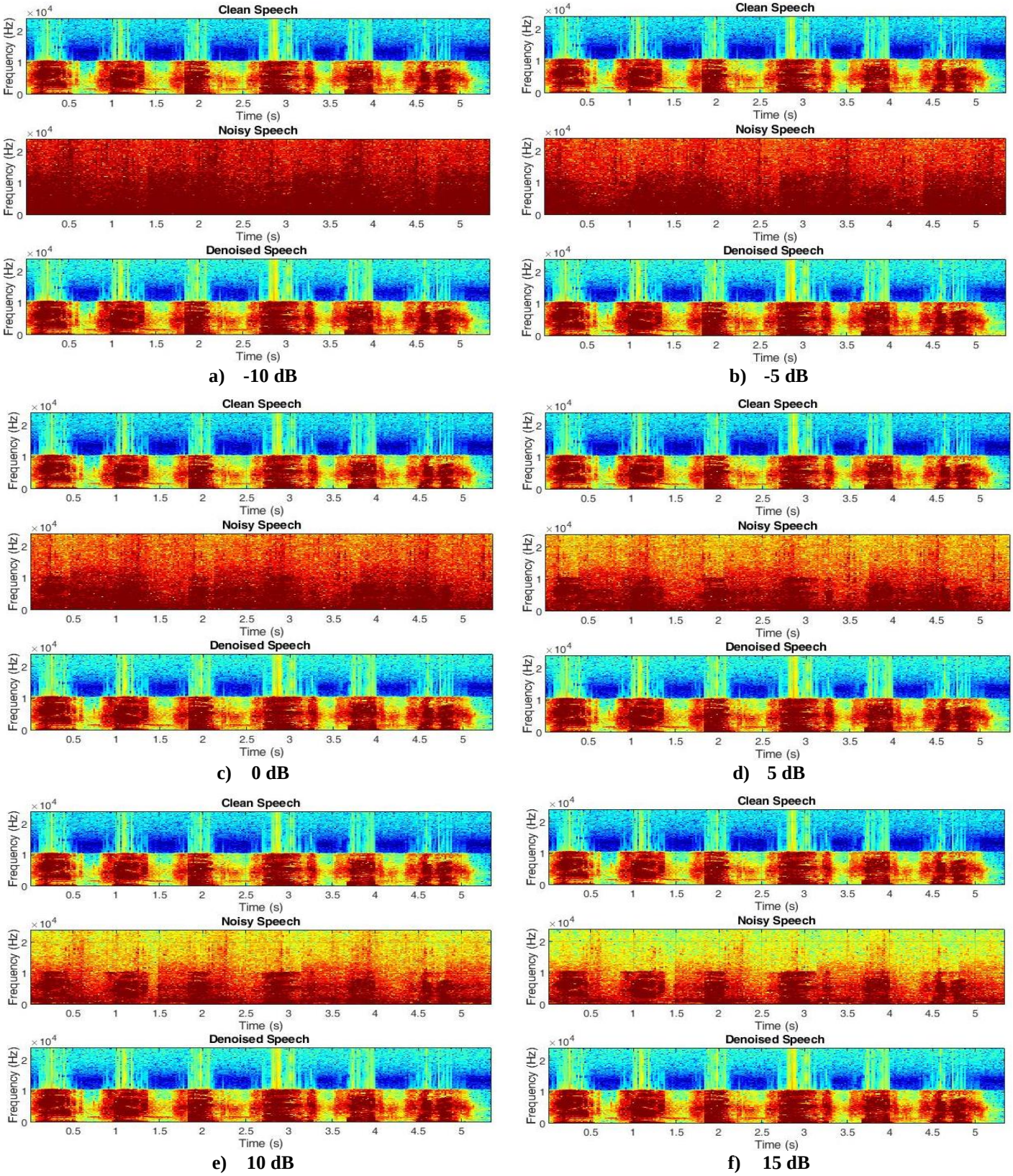
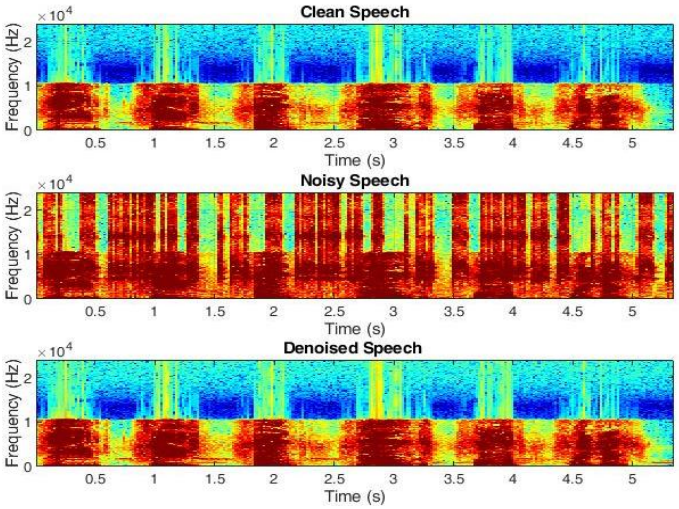
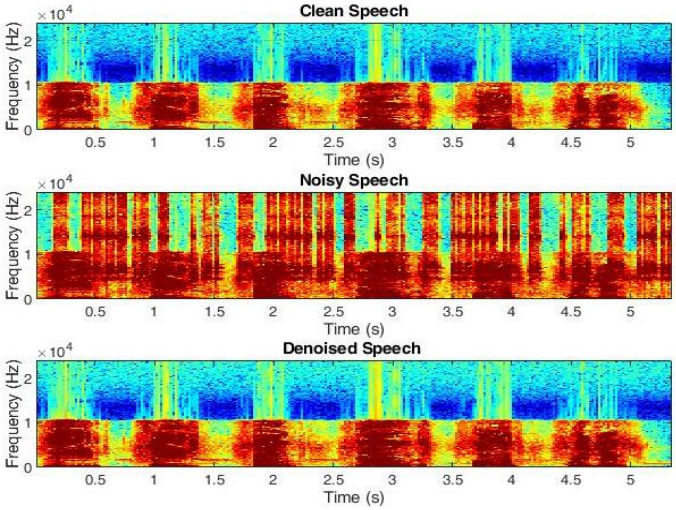


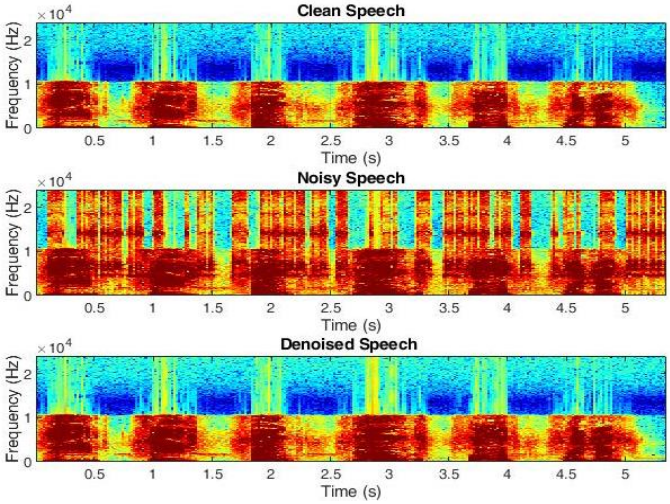
Figure 16 Deep CNN – Spectrogram Images of Street Noise for Alaryngeal Speech at various Noise Levels



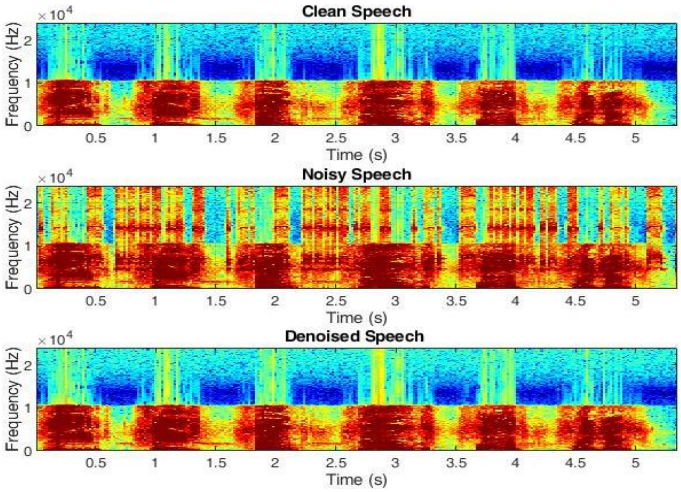
a) -10 dB



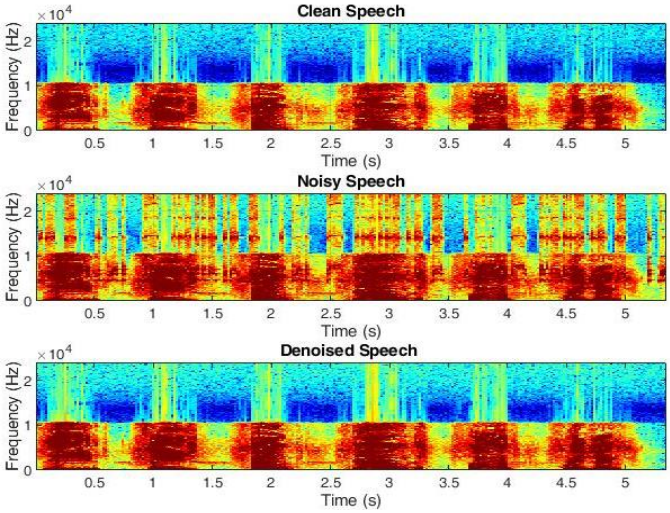
b) -5 dB



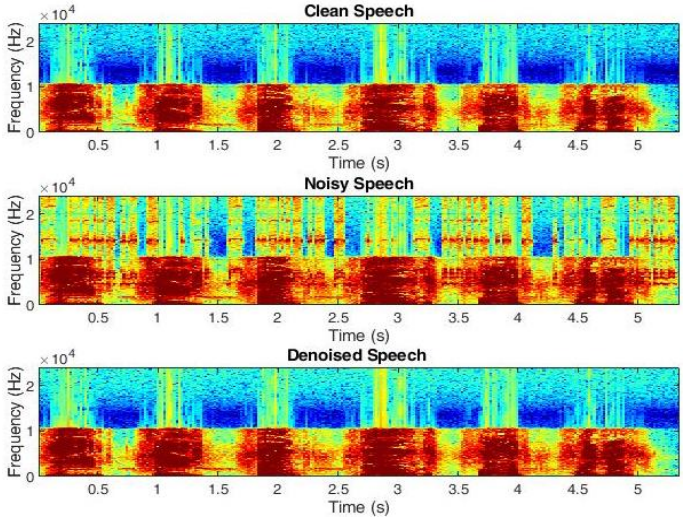
c) 0 dB



d) 5 dB



e) 10 dB



f) 15 dB

Figure 17 Deep CNN – Spectrogram Images of Train Whistle Noise for Alaryngeal Speech at various Noise Levels

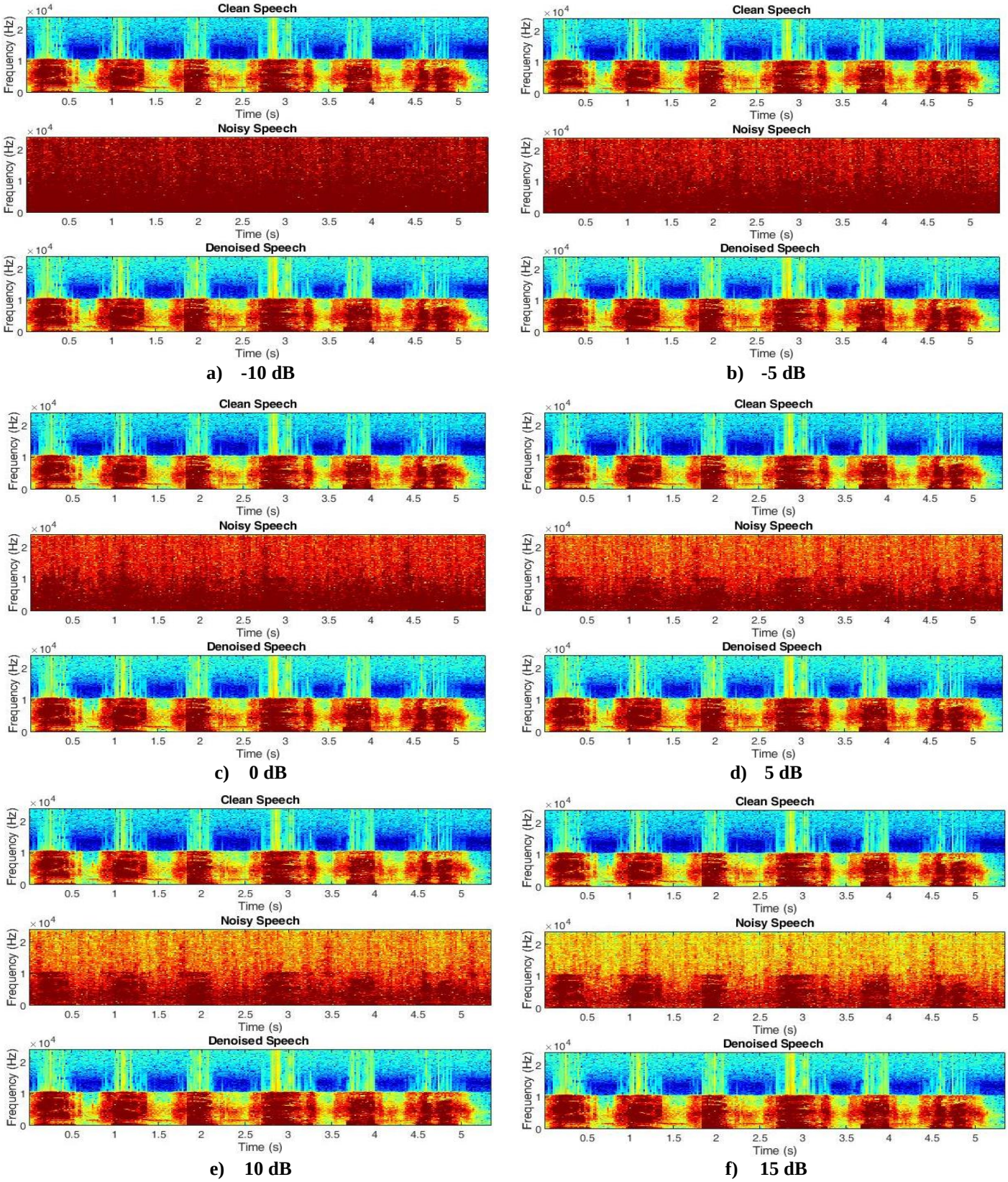
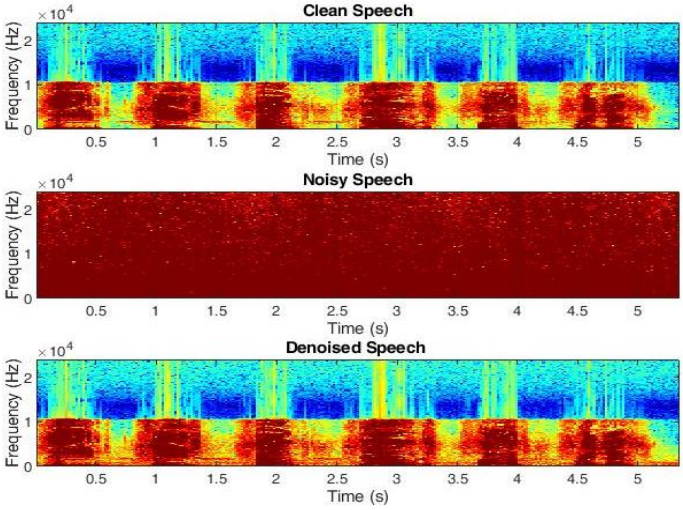
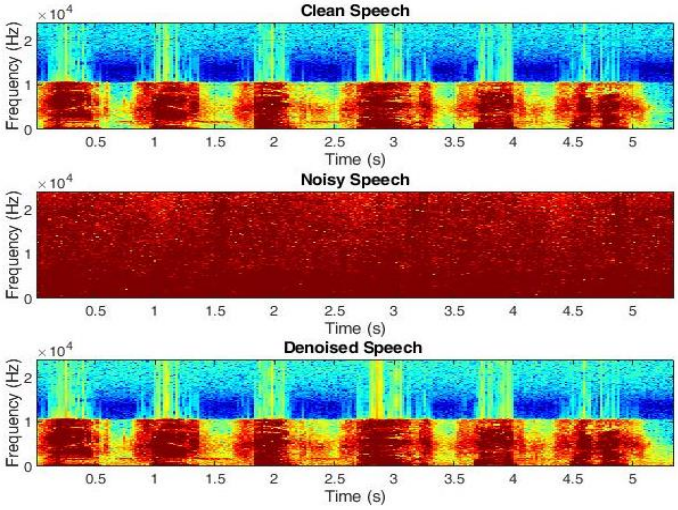


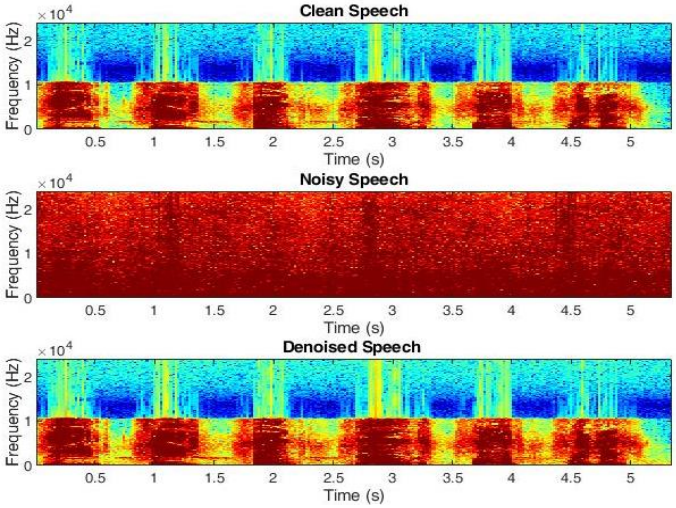
Figure 18 Deep CNN – Spectrogram Images of Restaurant Noise for Alaryngeal Speech at various Noise Levels



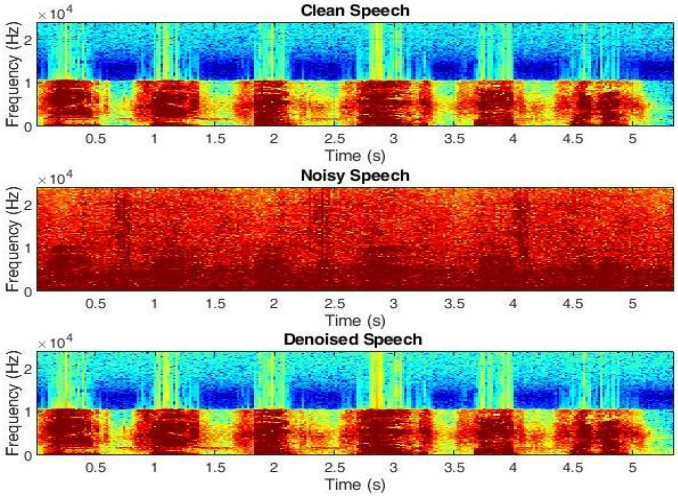
a) -10 dB



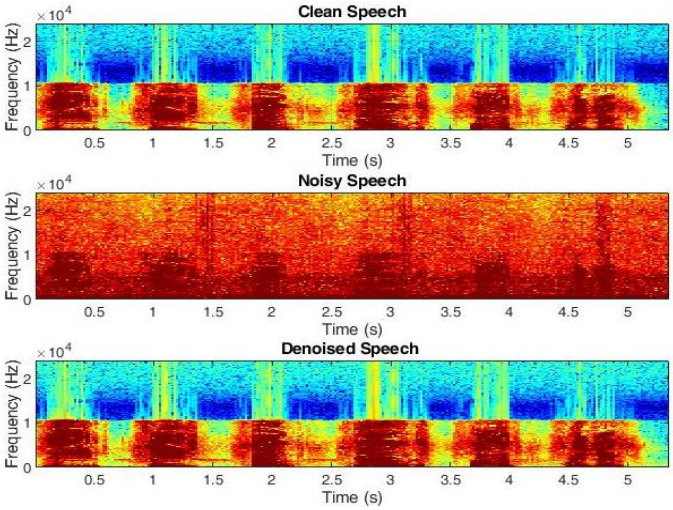
b) -5 dB



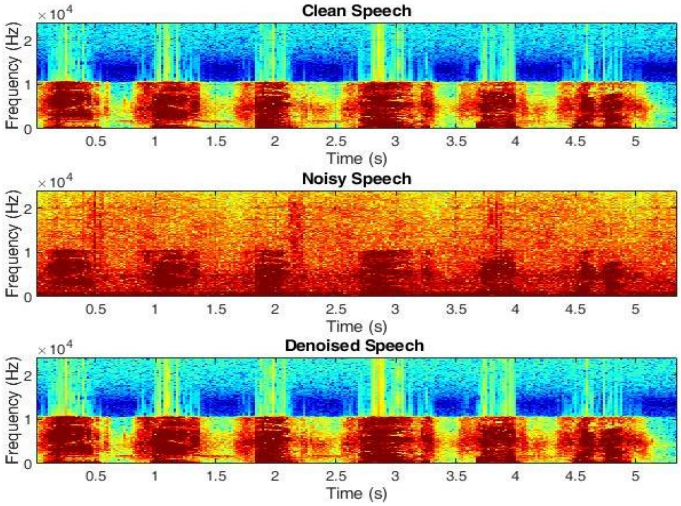
c) 0 dB



d) 5 dB

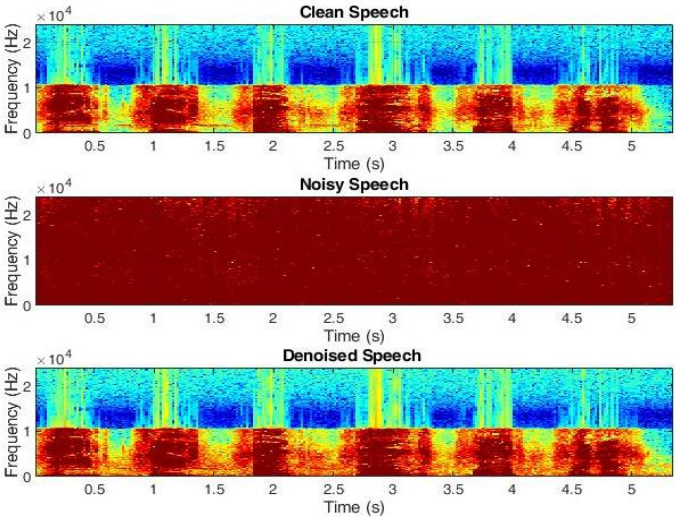


e) 10 dB

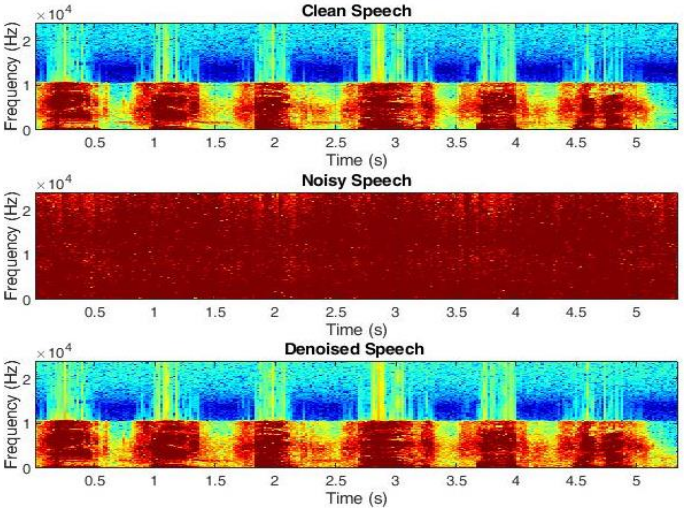


f) 15 dB

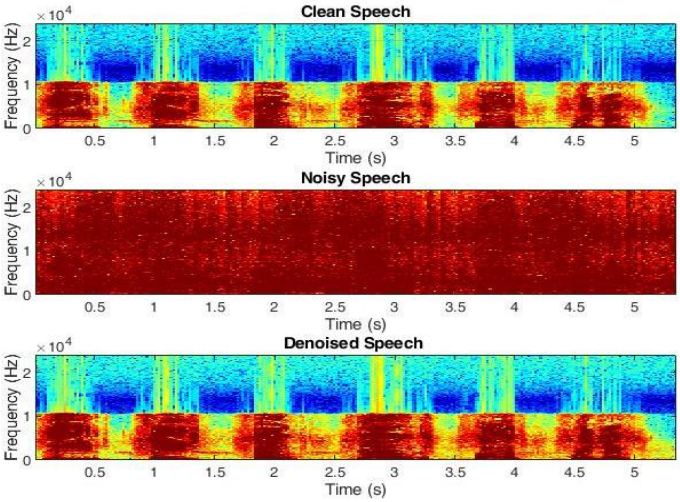
Figure 19 Deep CNN – Spectrogram Images of Car Noise for Alaryngeal Speech at various Noise Levels



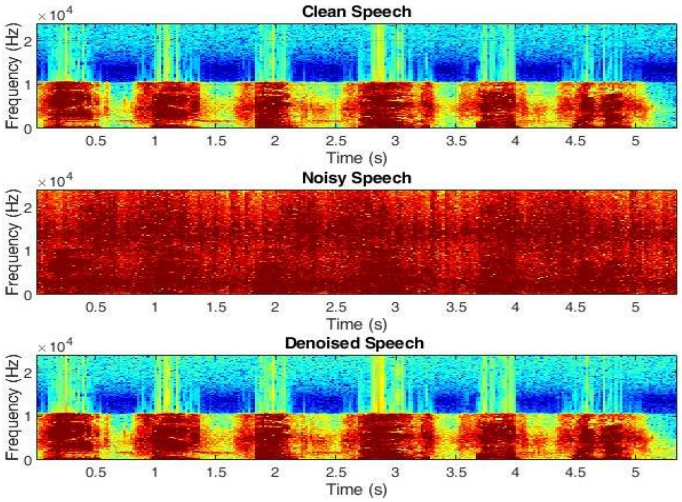
a) -10 dB



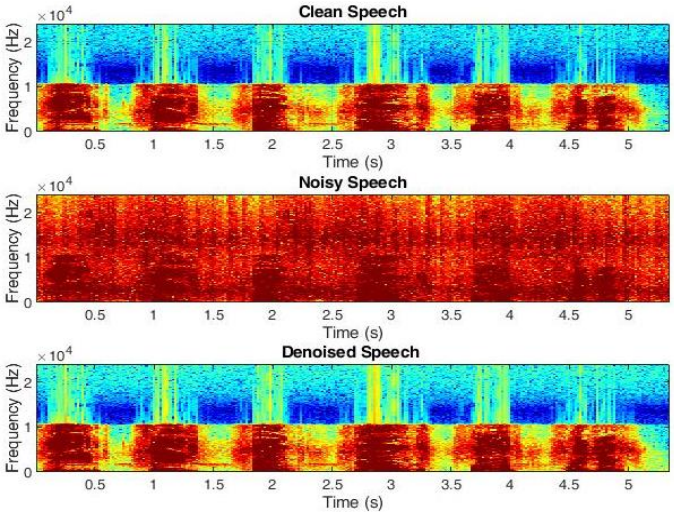
b) -5 dB



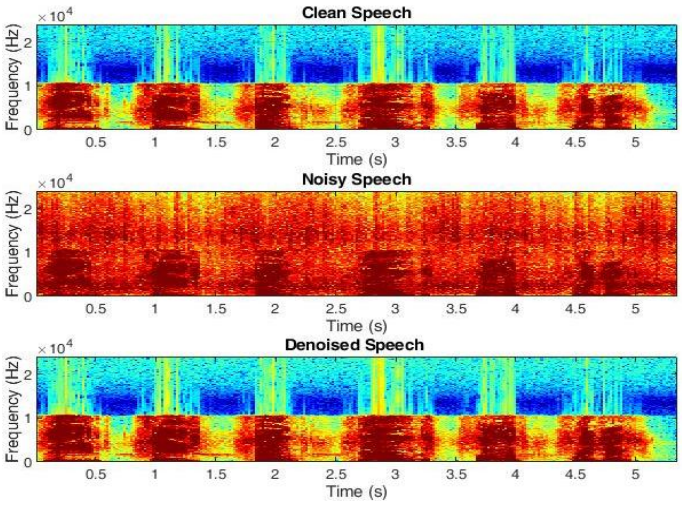
c) 0 dB



d) 5 dB



e) 10 dB



f) 15 dB

Figure 20 Deep CNN – Spectrogram Images of Subway Noise for Alaryngeal Speech at various Noise Levels

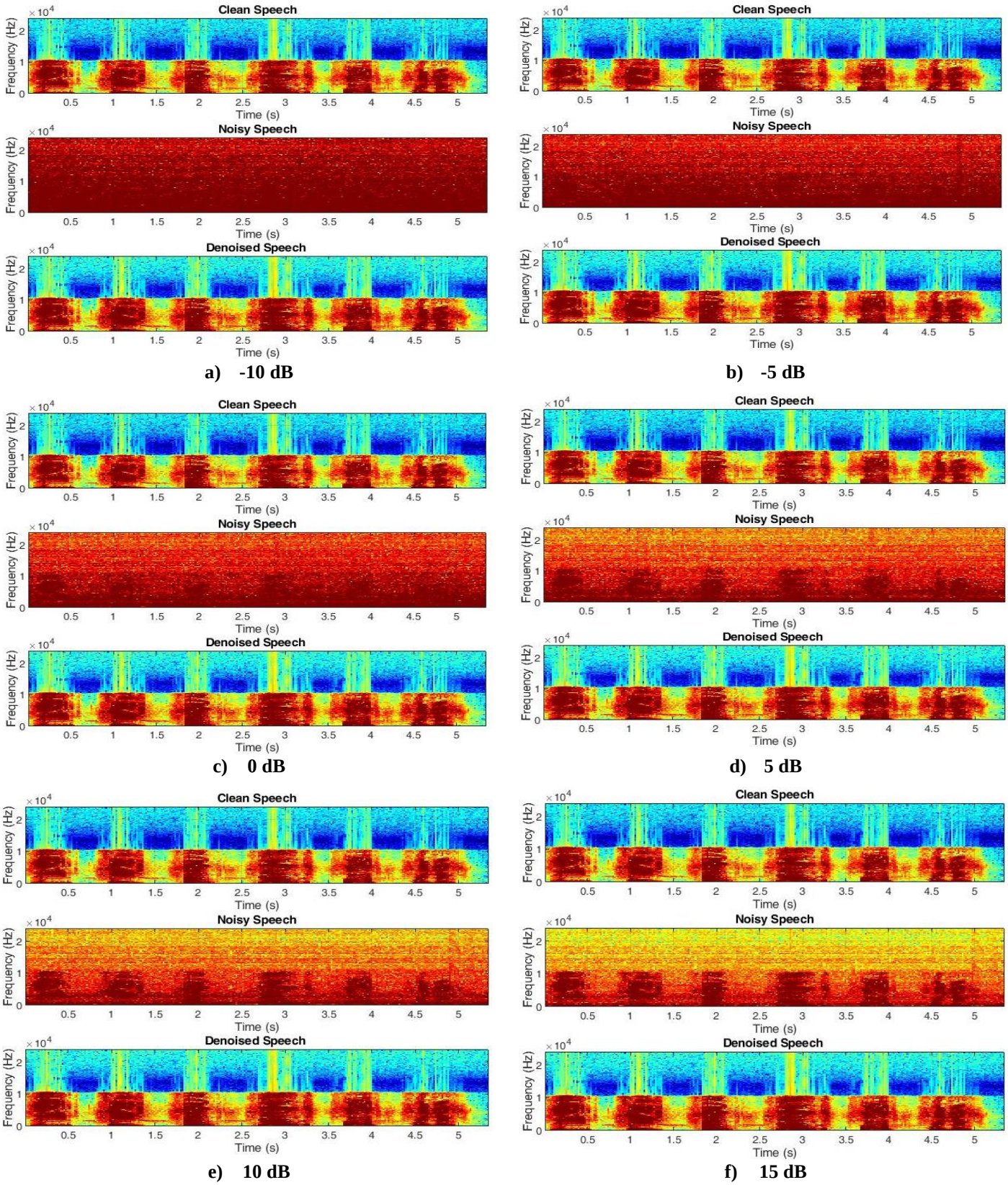
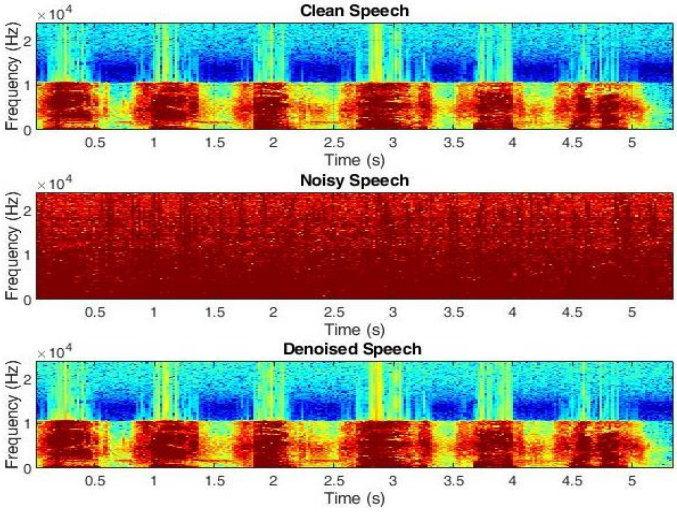
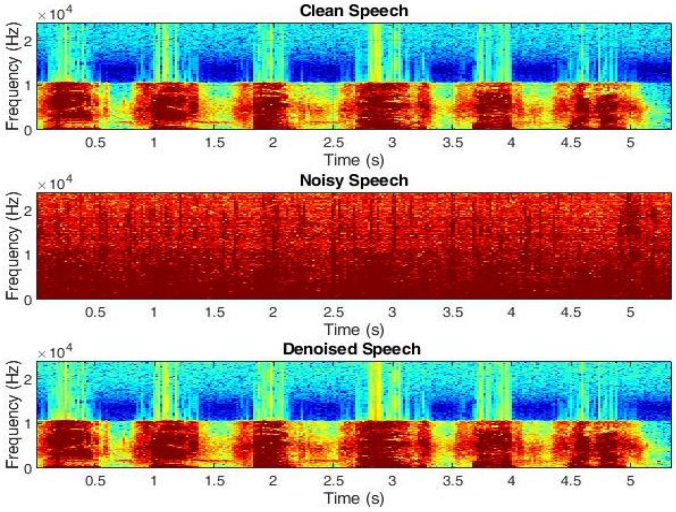


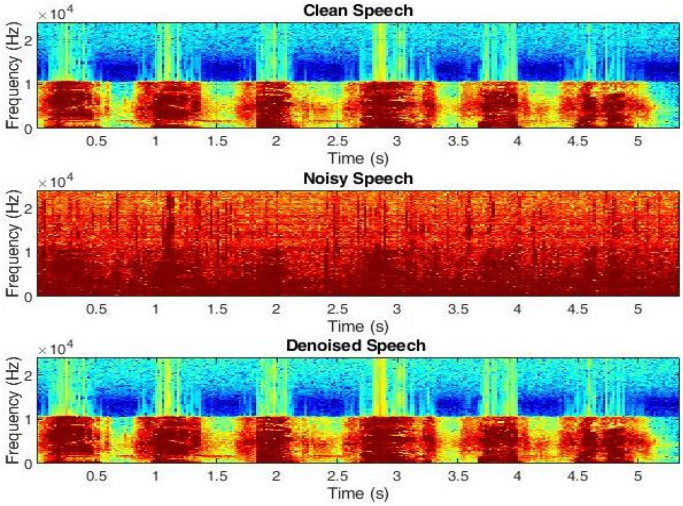
Figure 21 Modified LSTM – Spectrogram Images of Washing Machine Noise for Alaryngeal Speech at various Noise Levels



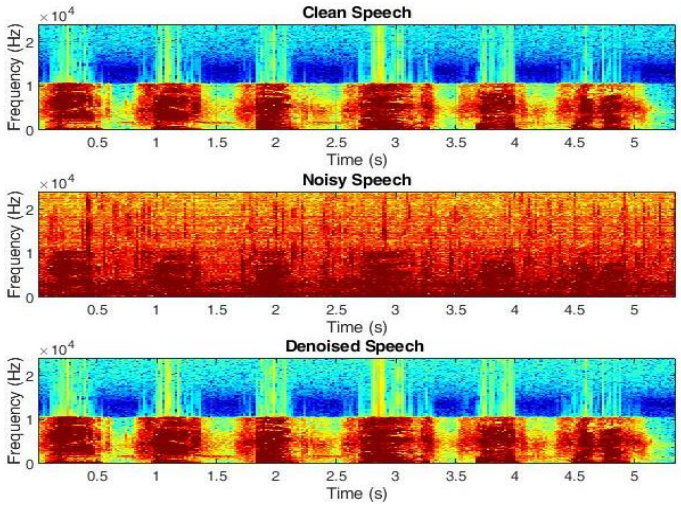
a) -10 dB



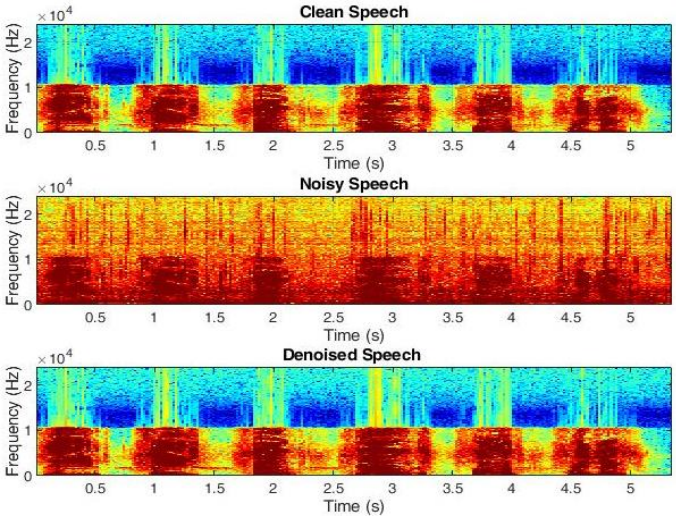
b) -5 dB



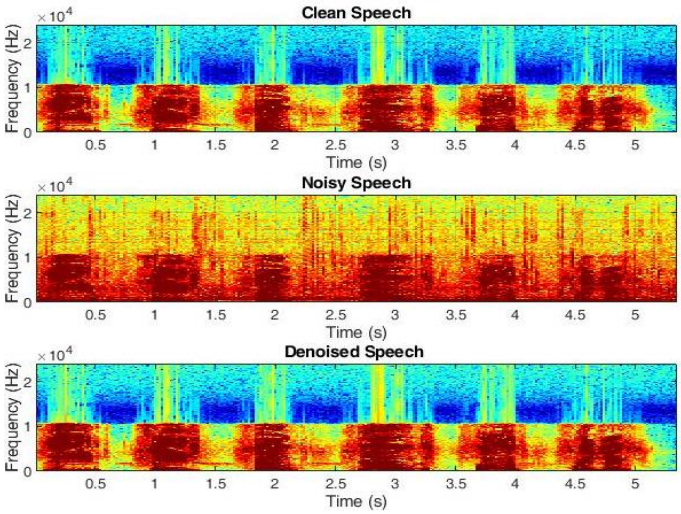
c) 0 dB



d) 5 dB



e) 10 dB



f) 15 dB

Figure 22 Modified LSTM – Spectrogram Images of Rainbow Noise for Alaryngeal Speech at various Noise Levels

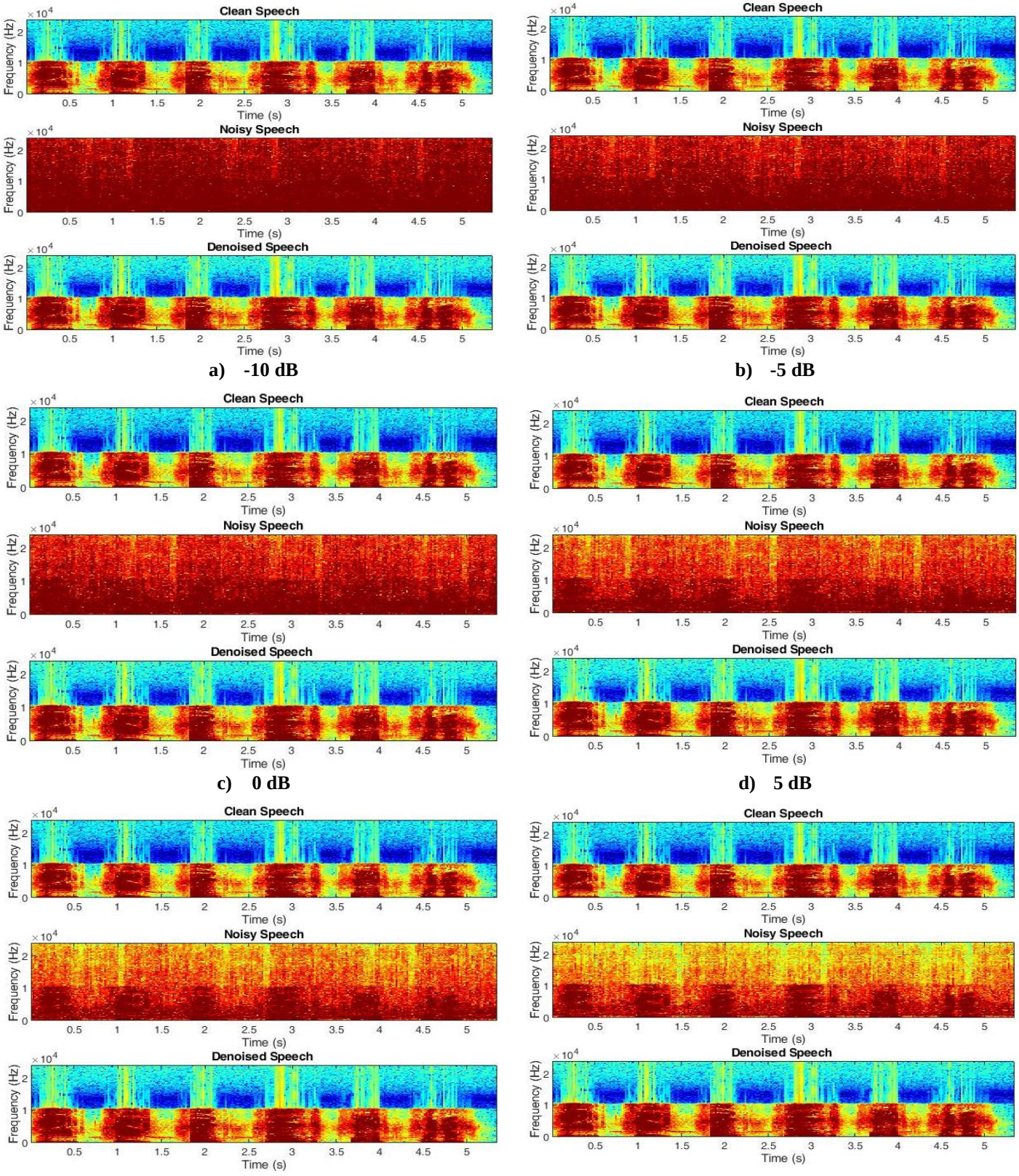
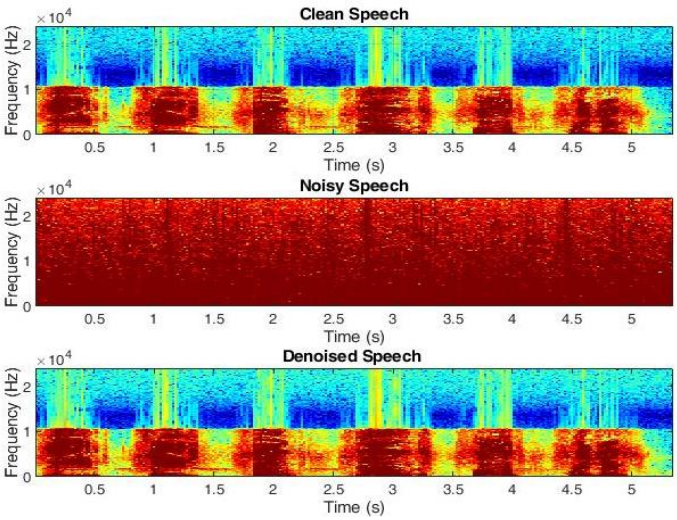
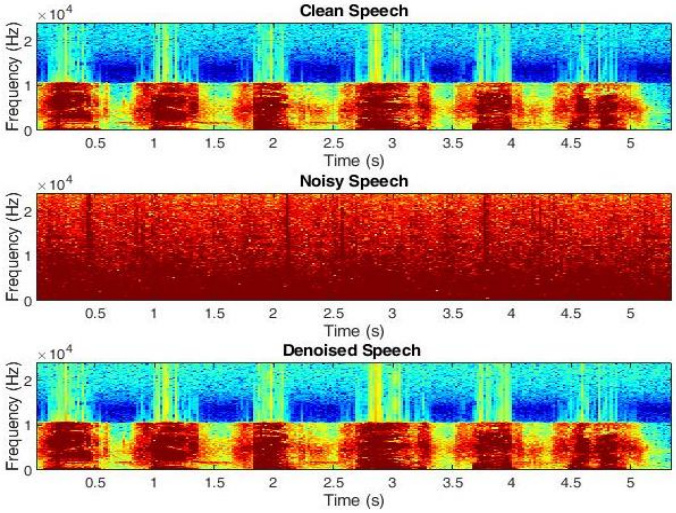


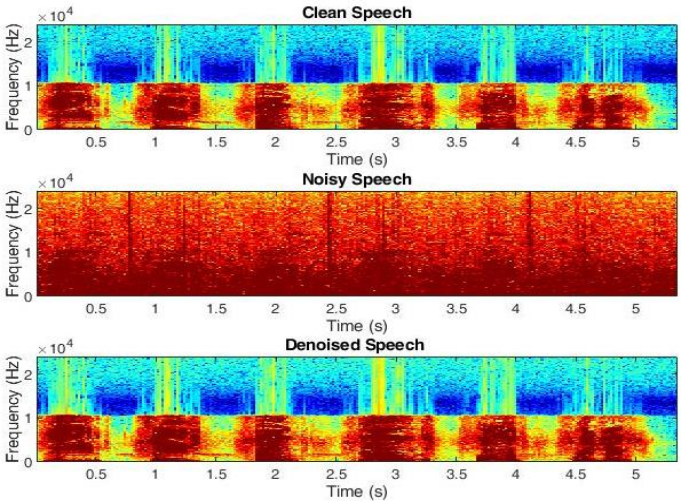
Figure 23 Modified LSTM – Spectrogram Images of Babble Noise for Alaryngeal Speech at various Noise Levels



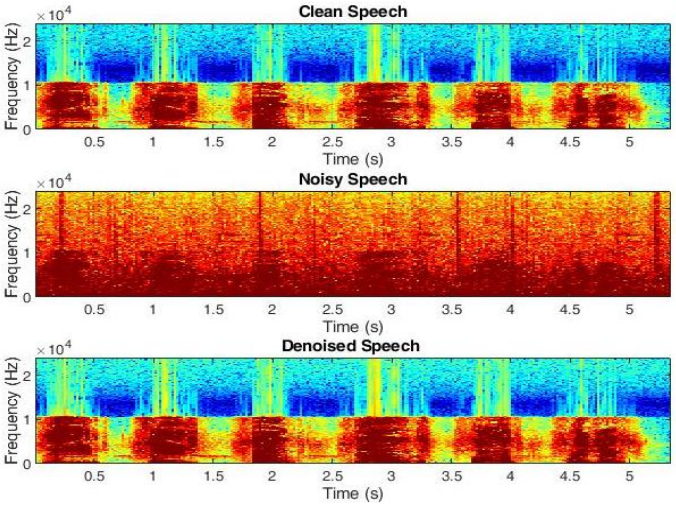
a) -10 dB



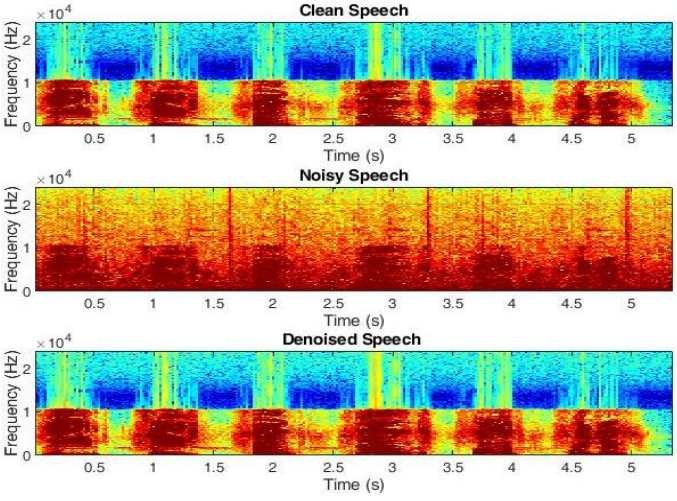
b) -5 dB



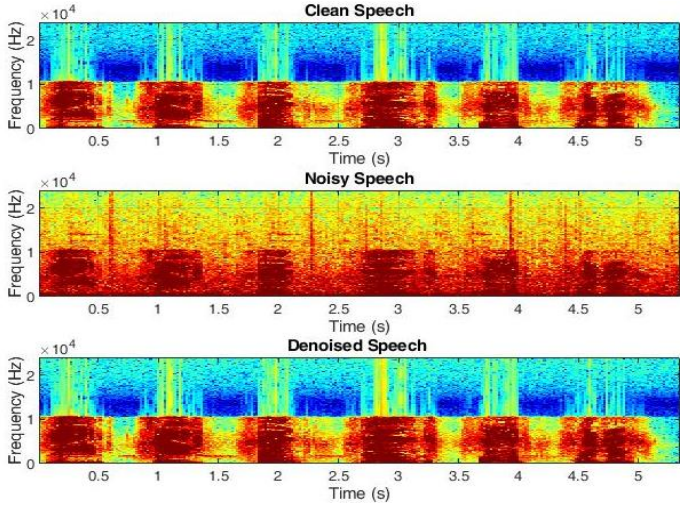
c) 0 dB



d) 5 dB

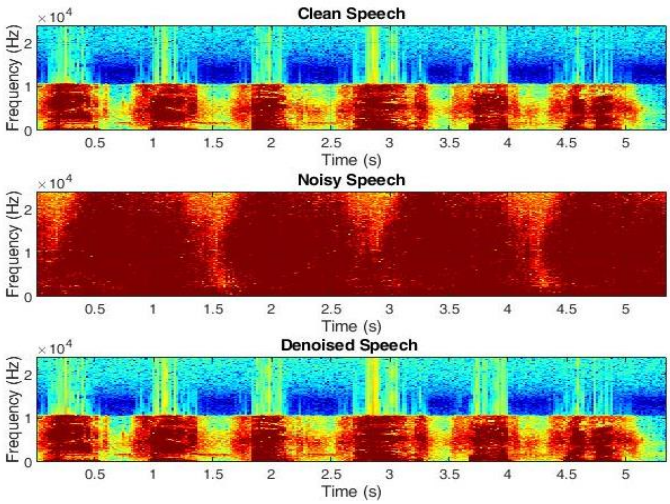


e) 10 dB

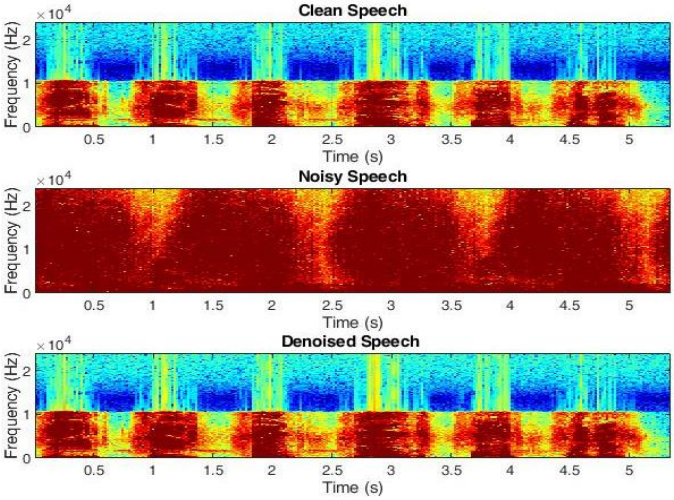


f) 15 dB

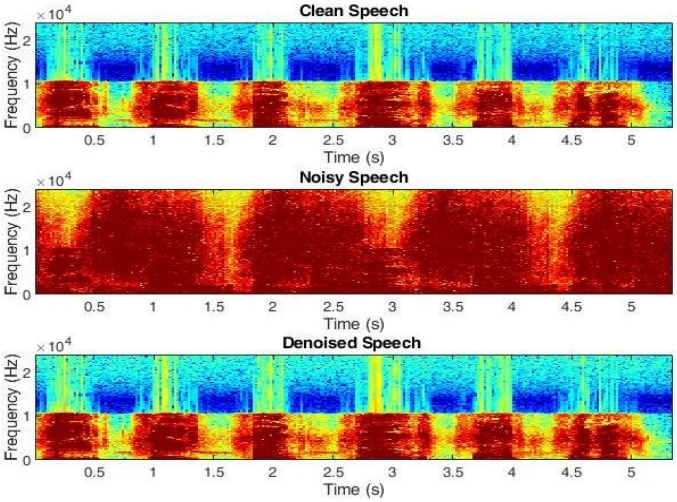
Figure 24 Modified LSTM – Spectrogram Images of Airport Noise for Alaryngeal Speech at various Noise Levels



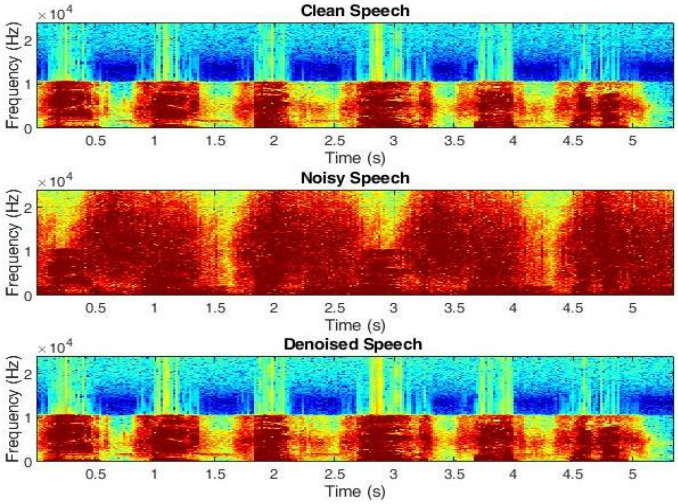
a) -10 dB



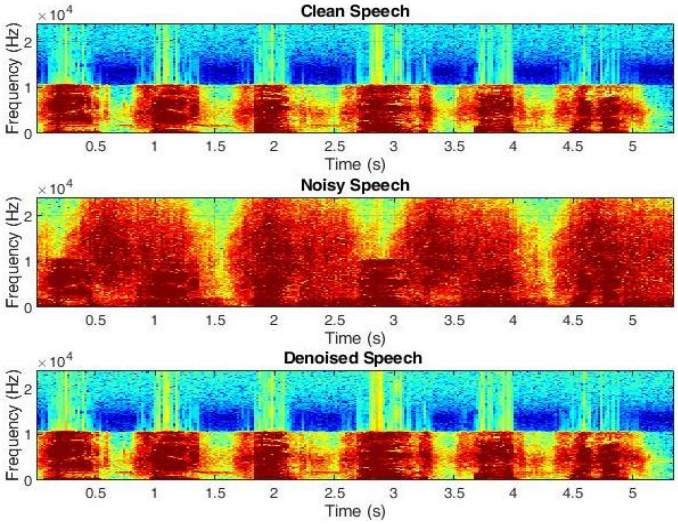
b) -5 dB



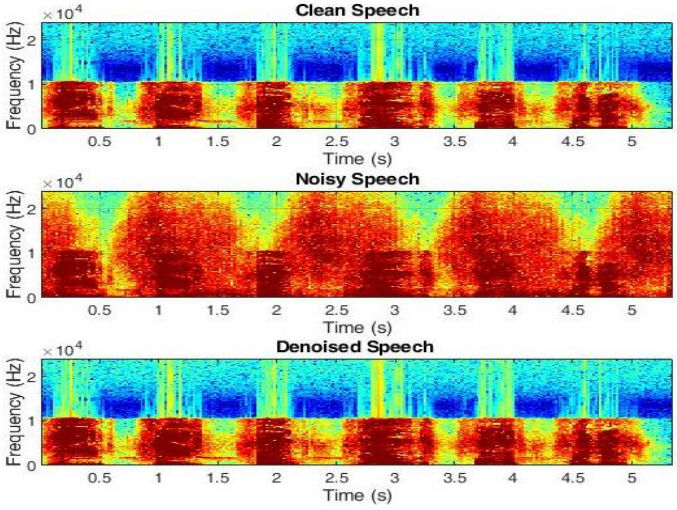
c) 0 dB



d) 5 dB

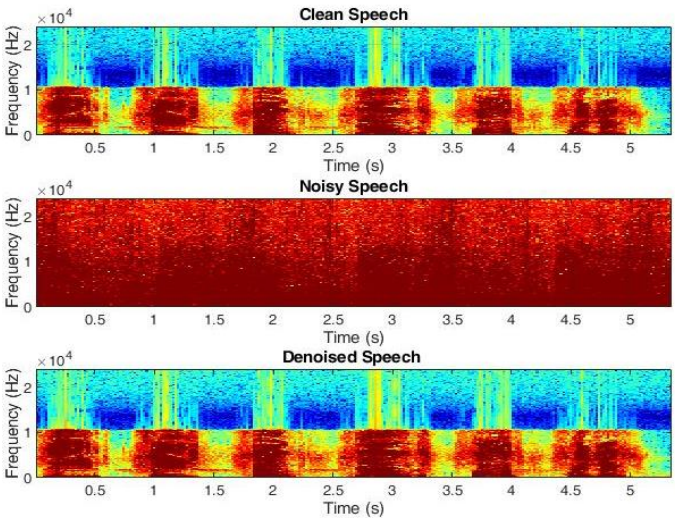


e) 10 dB

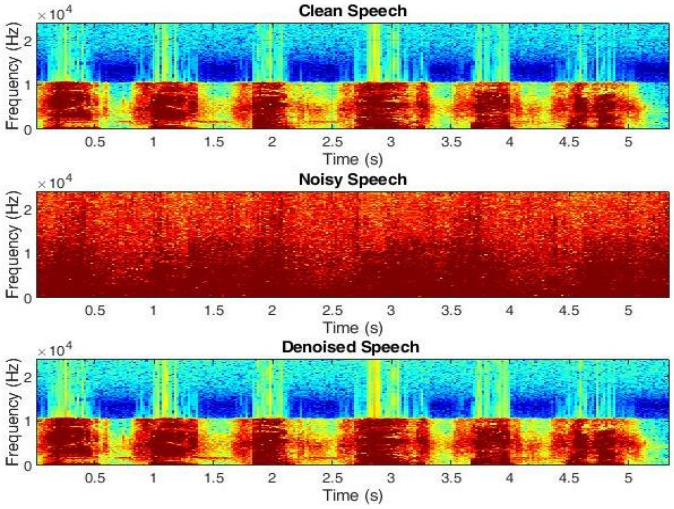


f) 15 dB

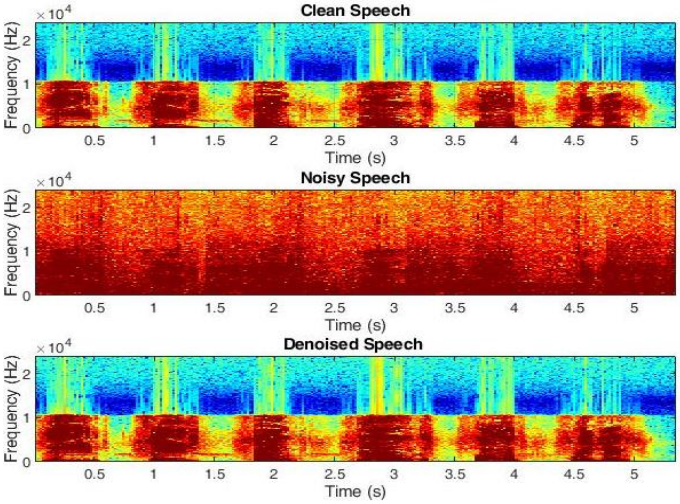
Figure 25 Modified LSTM – Spectrogram Images of Jet plane Noise for Alaryngeal Speech at various Noise Levels



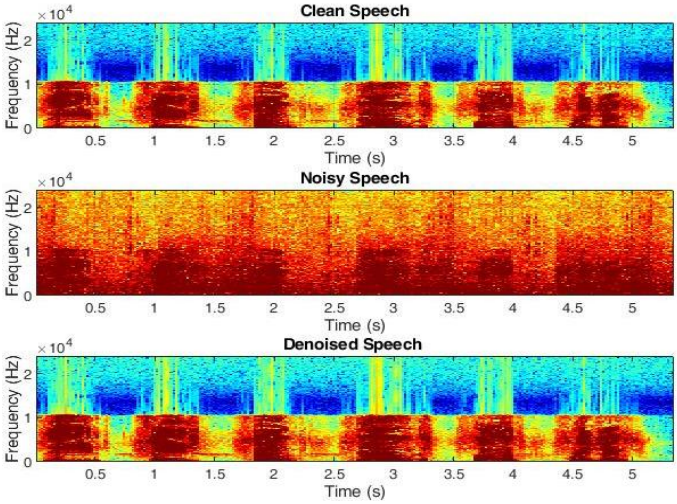
a) -10 dB



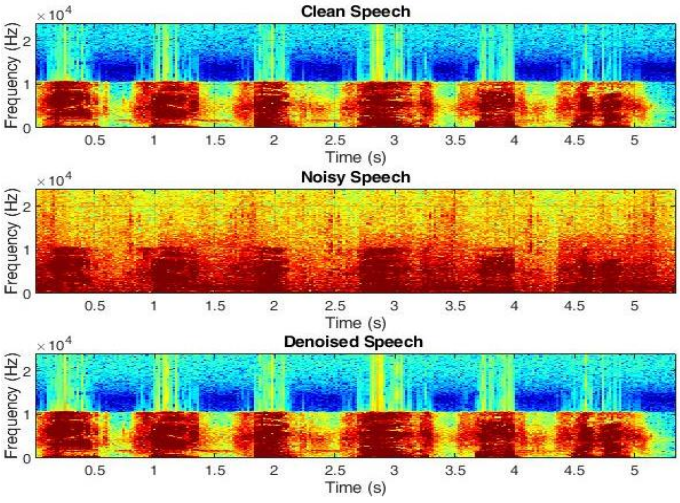
b) -5 dB



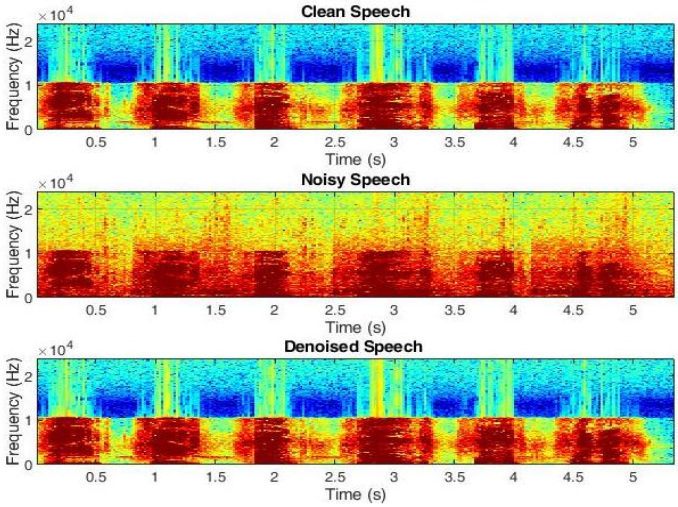
c) 0 dB



d) 5 dB

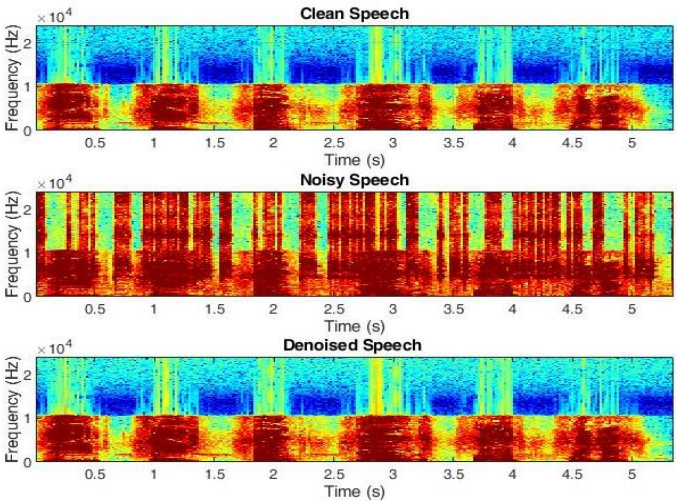


e) 10 dB

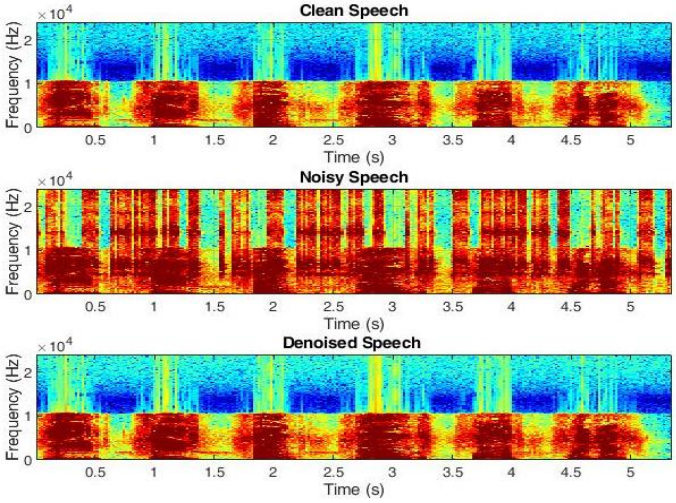


f) 15 dB

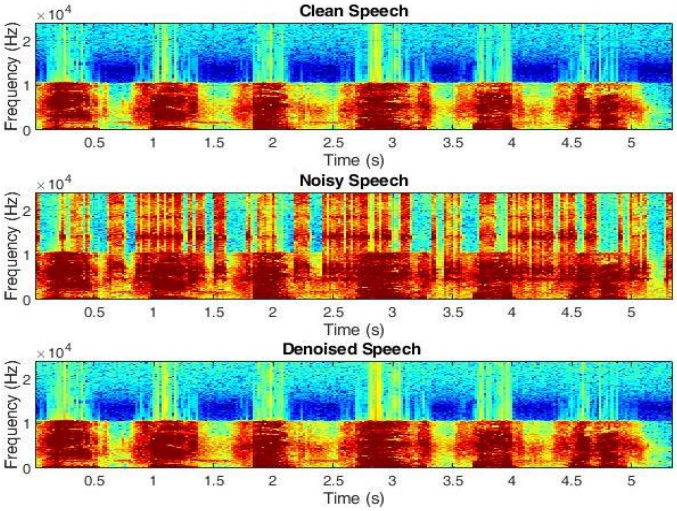
Figure 26 Modified LSTM – Spectrogram Images of Street Noise for Alaryngeal Speech at various Noise Levels



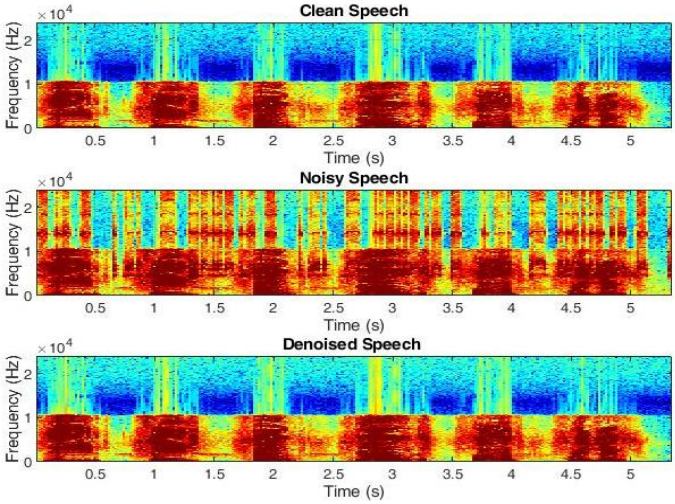
a) -10 dB



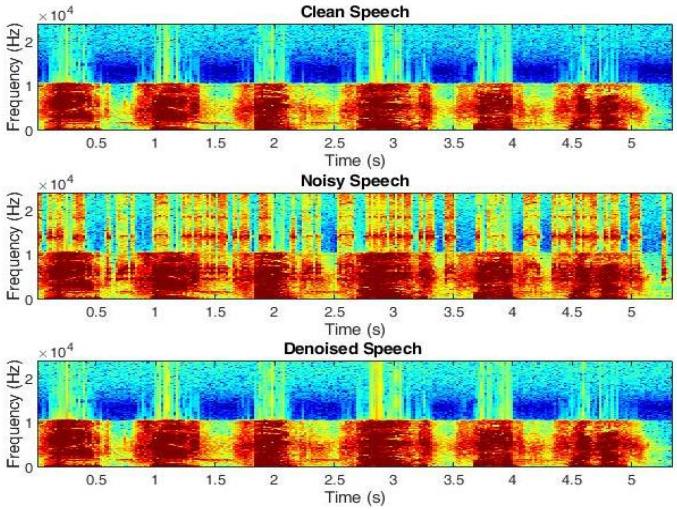
b) -5 dB



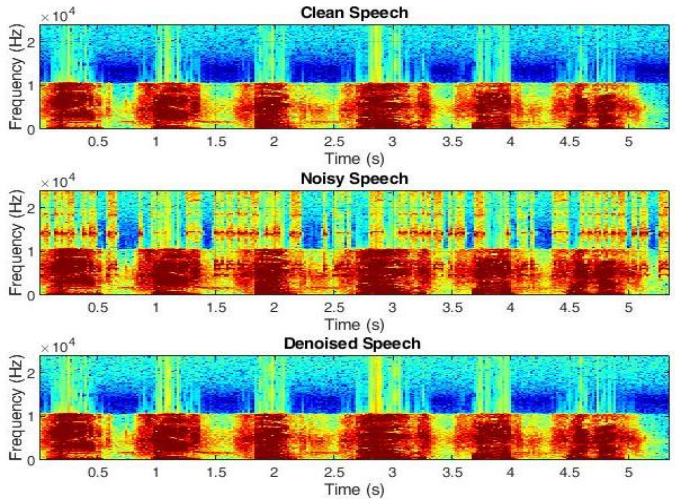
c) 0 dB



d) 5 dB

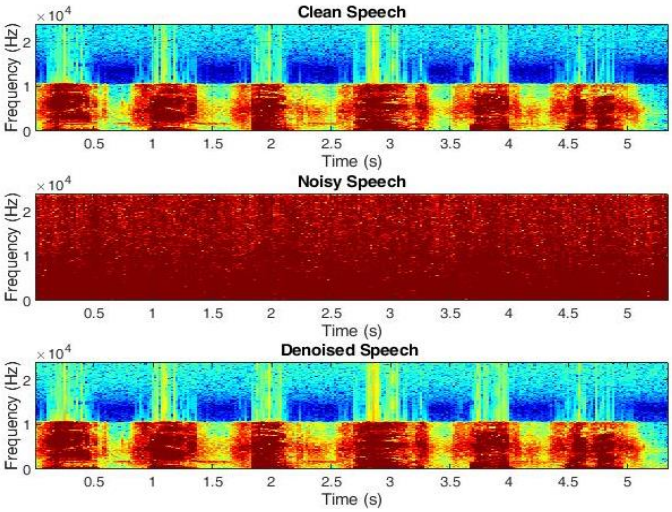


e) 10 dB

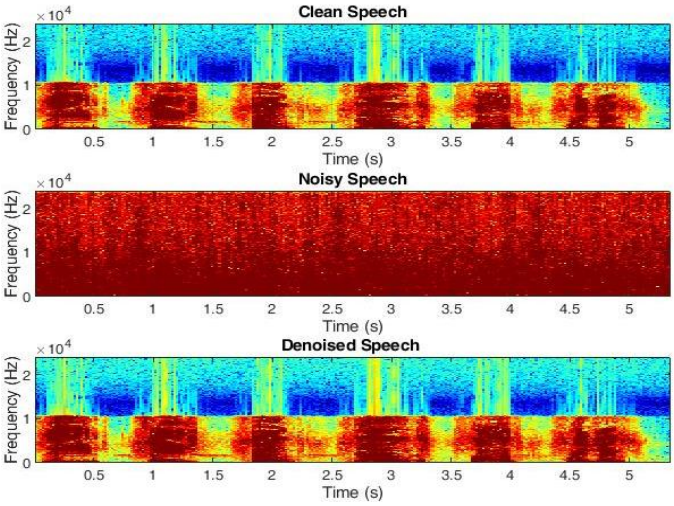


f) 15 dB

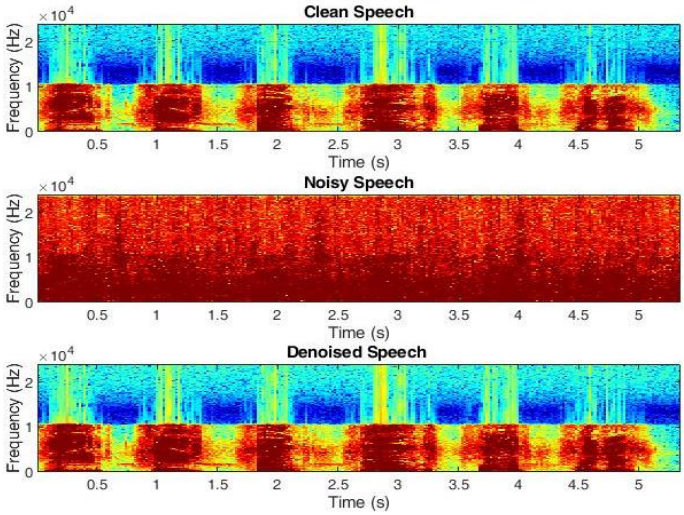
Figure 27 Modified LSTM – Spectrogram Images of Train Whistle Noise for Alaryngeal Speech at various Noise Levels



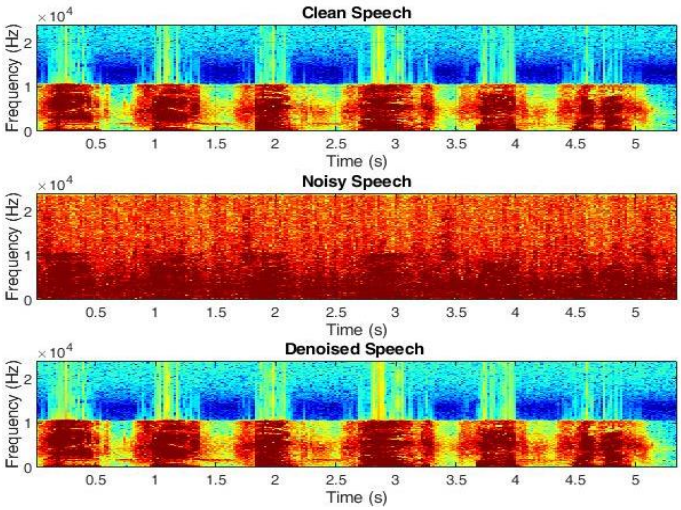
a) -10 dB



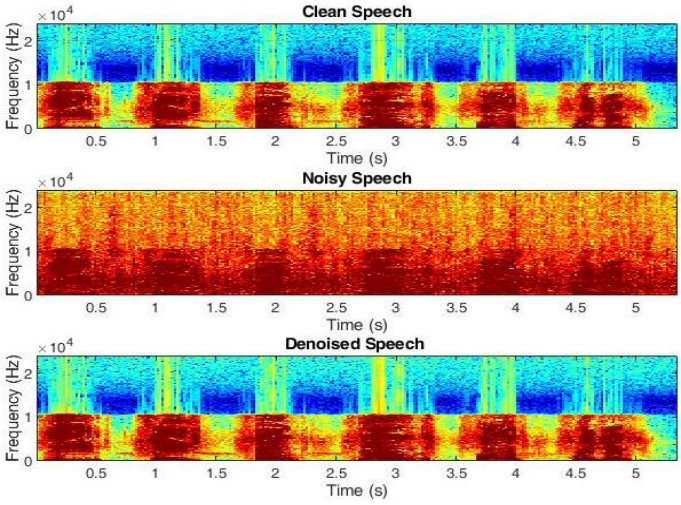
b) -5 dB



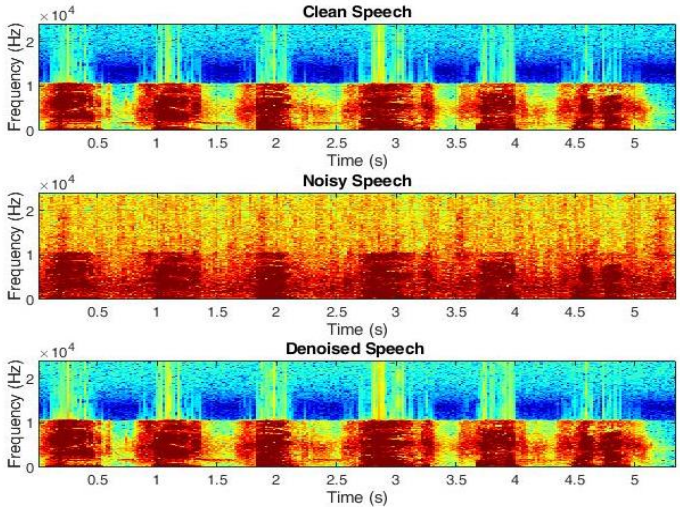
c) 0 dB



d) 5 dB



e) 10 dB



f) 15 dB

Figure 28 Modified LSTM – Spectrogram Images of Restaurant Noise for Alaryngeal Speech at various Noise Levels

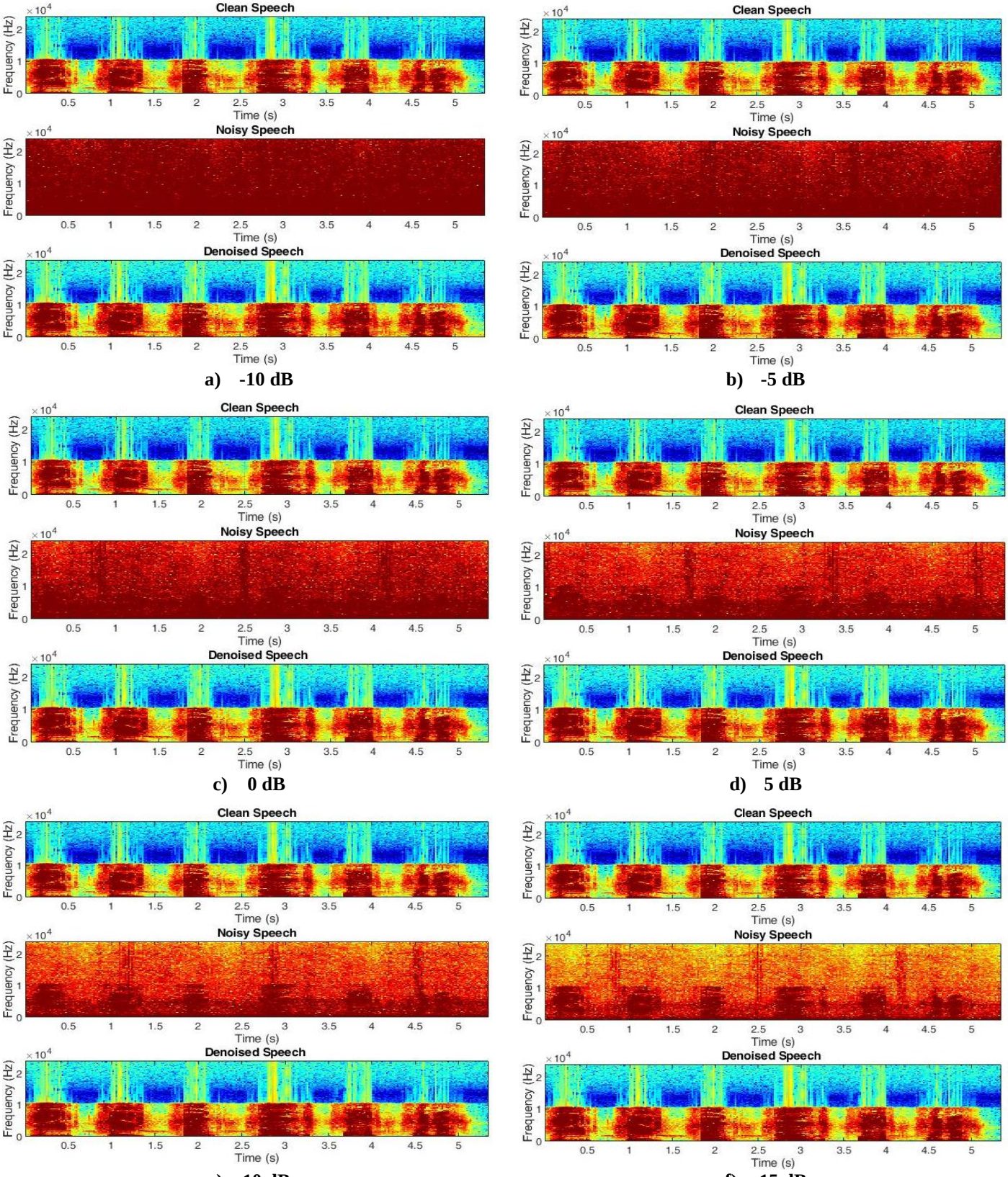


Figure 29 Modified LSTM – Spectrogram Images of Car Noise for Alaryngeal Speech at various Noise Levels

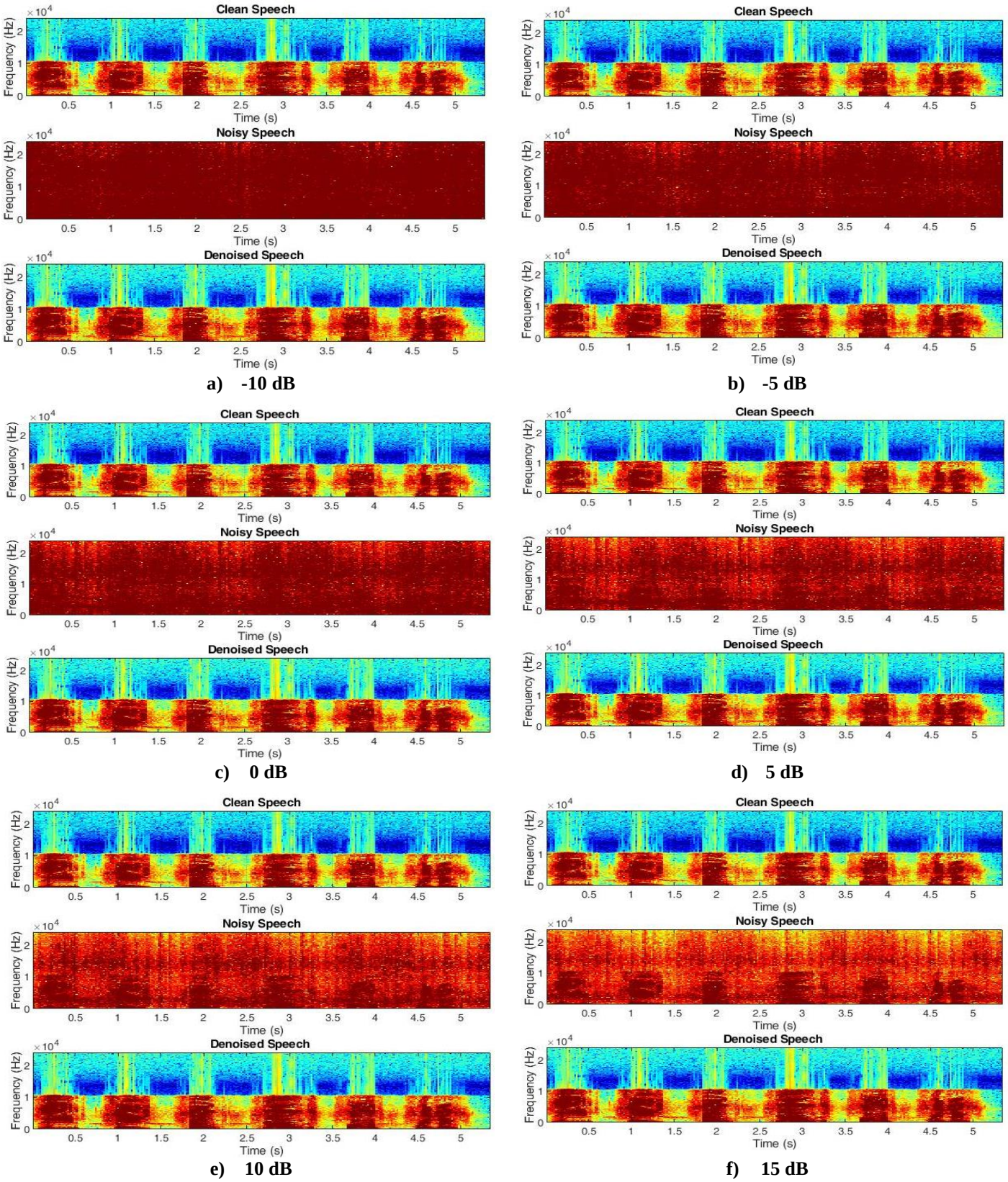
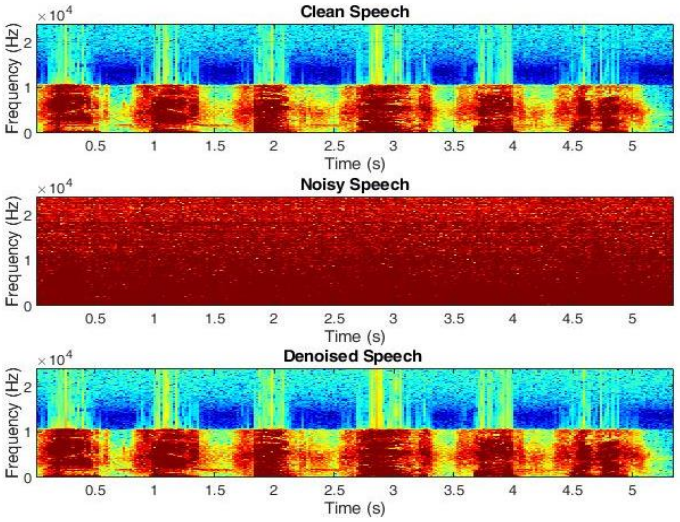
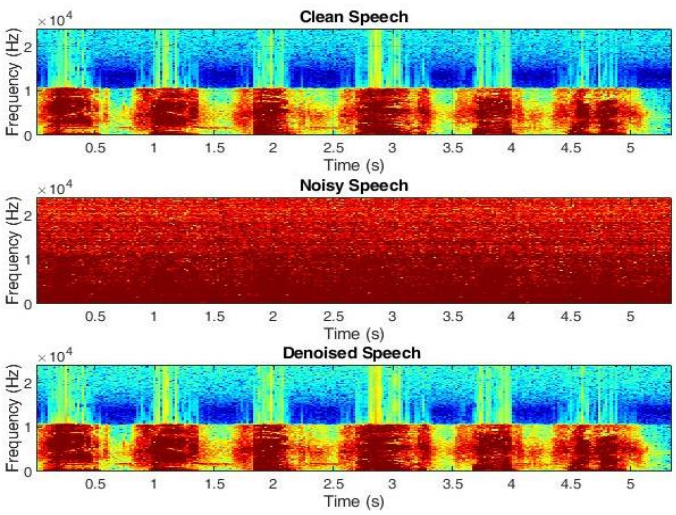


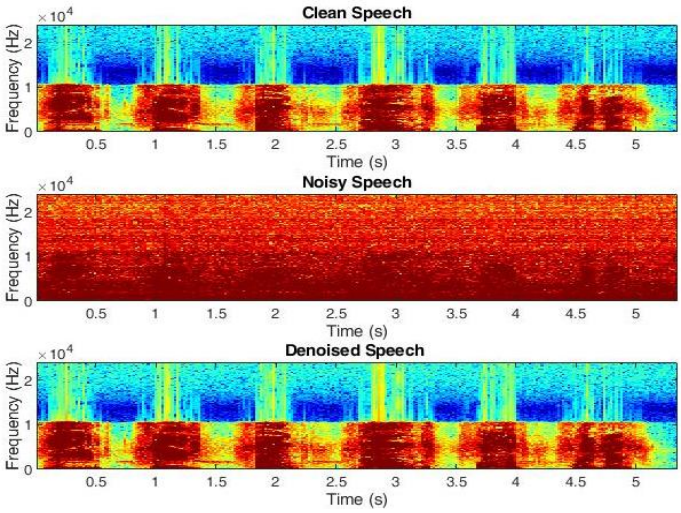
Figure 30 Modified LSTM – Spectrogram Images of Subway Noise for Alaryngeal Speech at various Noise Levels



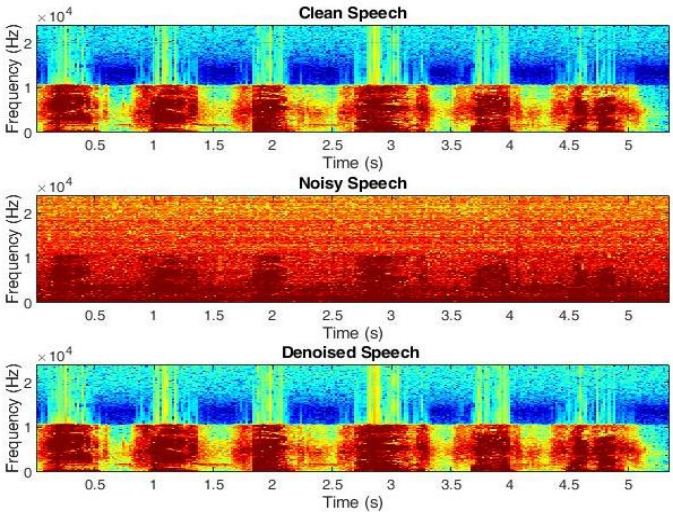
a) -10 dB



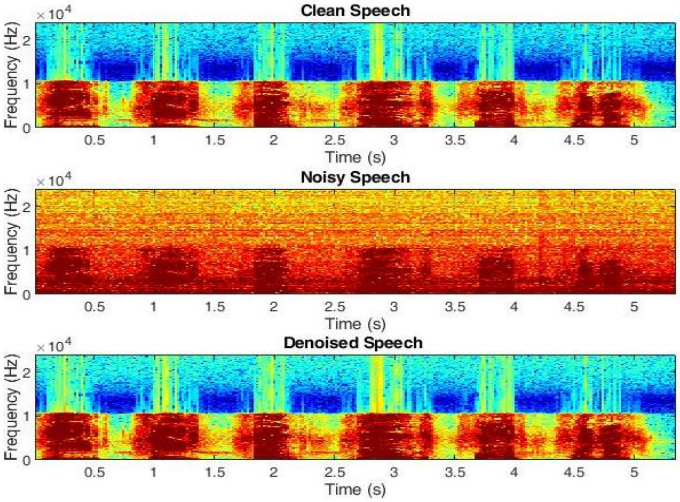
b) -5 dB



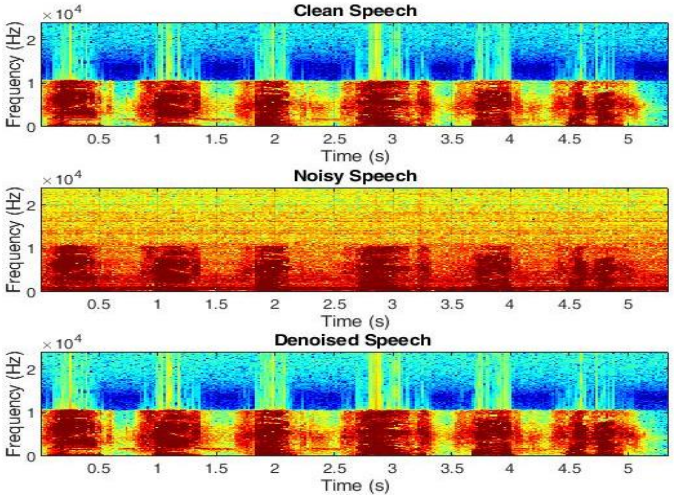
c) 0 dB



d) 5 dB



e) 10 dB



f) 15 dB

Figure 31 Modified FCRN – Spectrogram Images of Washing Machine Noise for Alaryngeal Speech at various Noise Levels

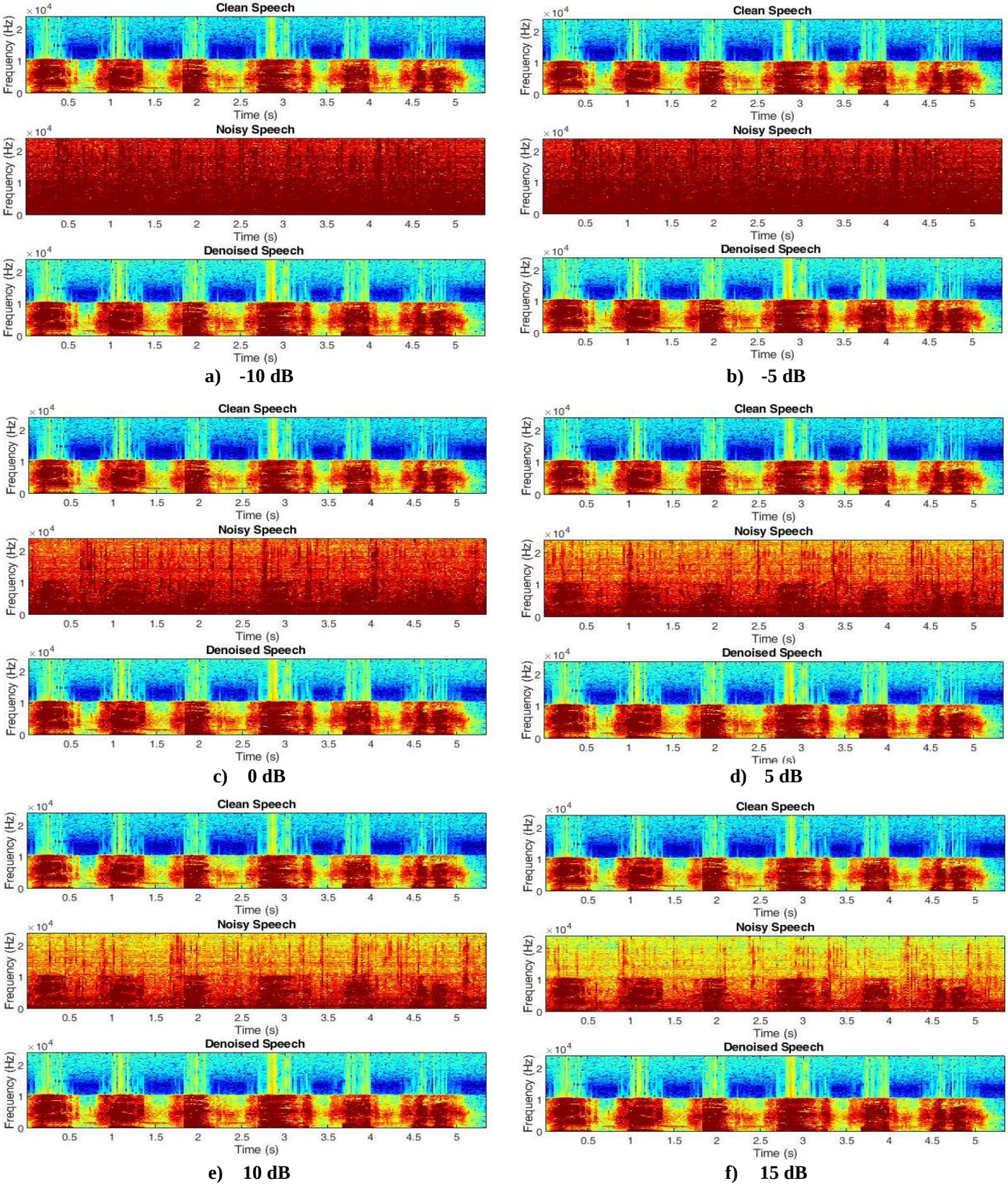
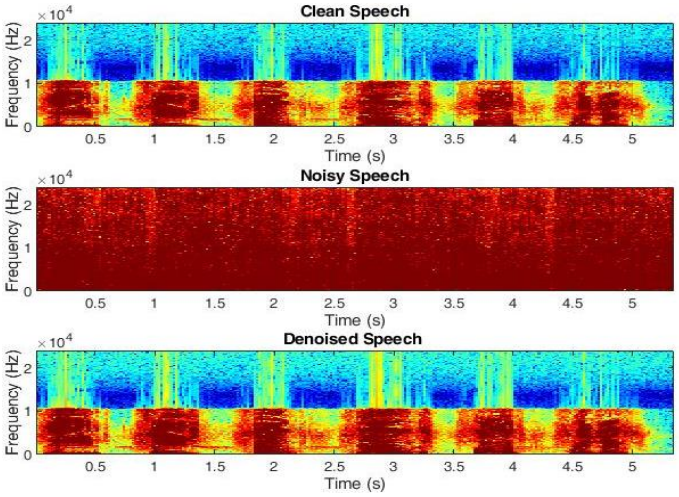
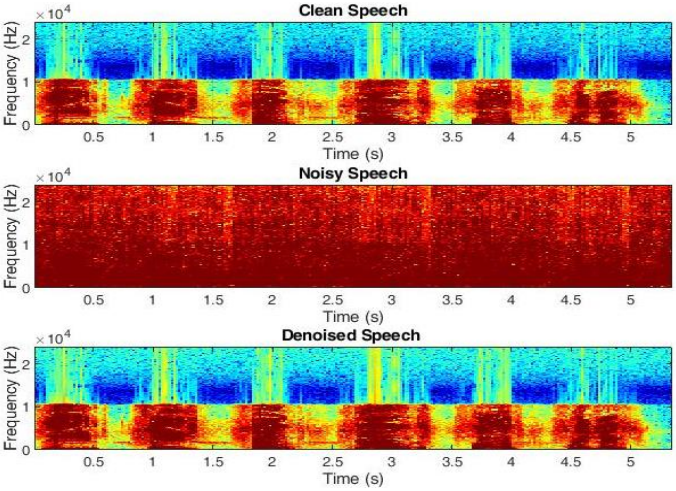


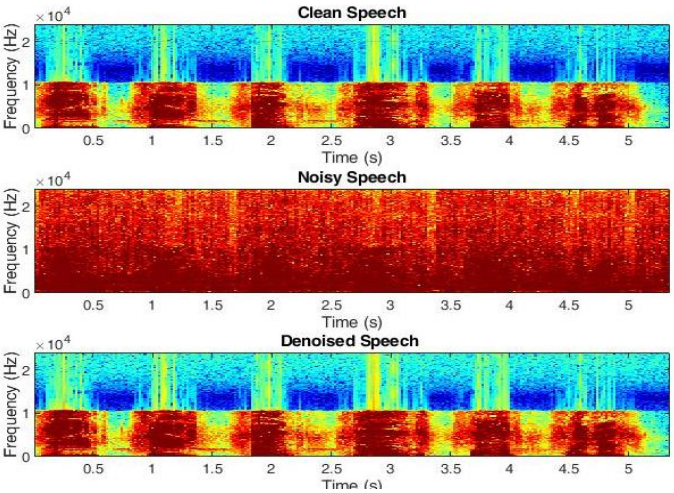
Figure 32 Modified FCRN – Spectrogram Images of Rainbow Noise for Alaryngeal Speech at various Noise Levels



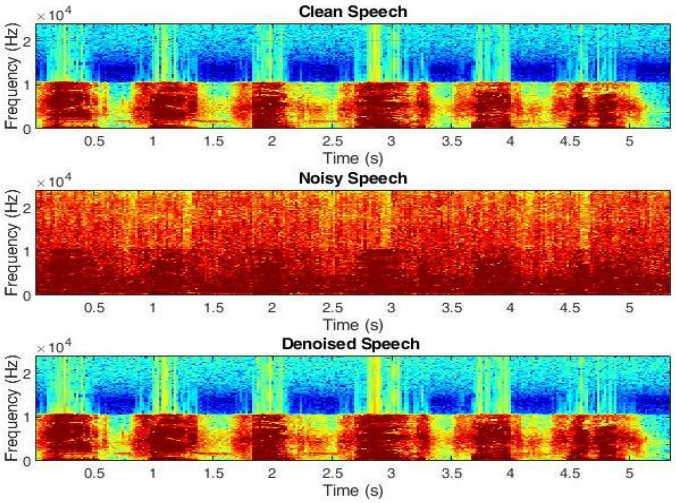
a) -10 dB



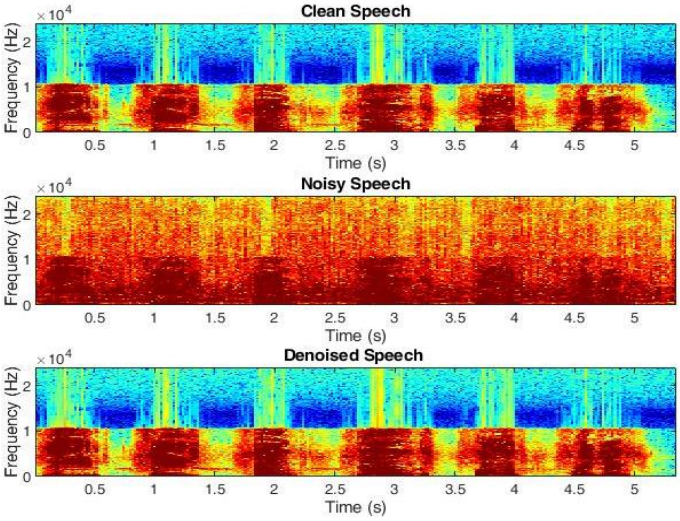
b) -5 dB



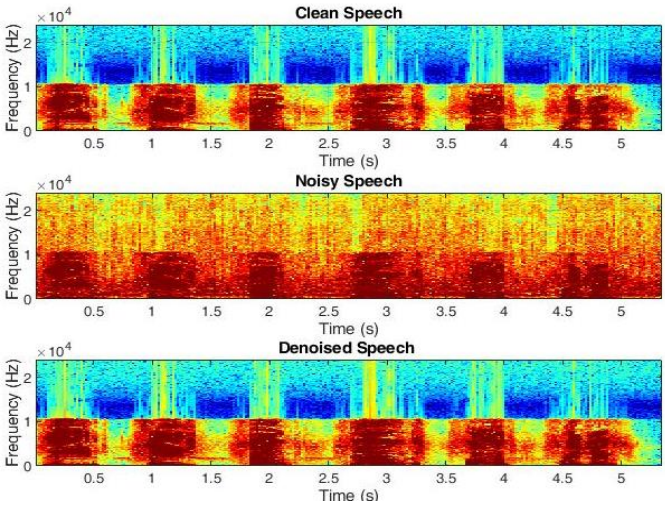
c) 0 dB



d) 5 dB



e) 10 dB



f) 15 dB

Figure 33 Modified FCRN – Spectrogram Images of Babble Noise for Alaryngeal Speech at various Noise Levels

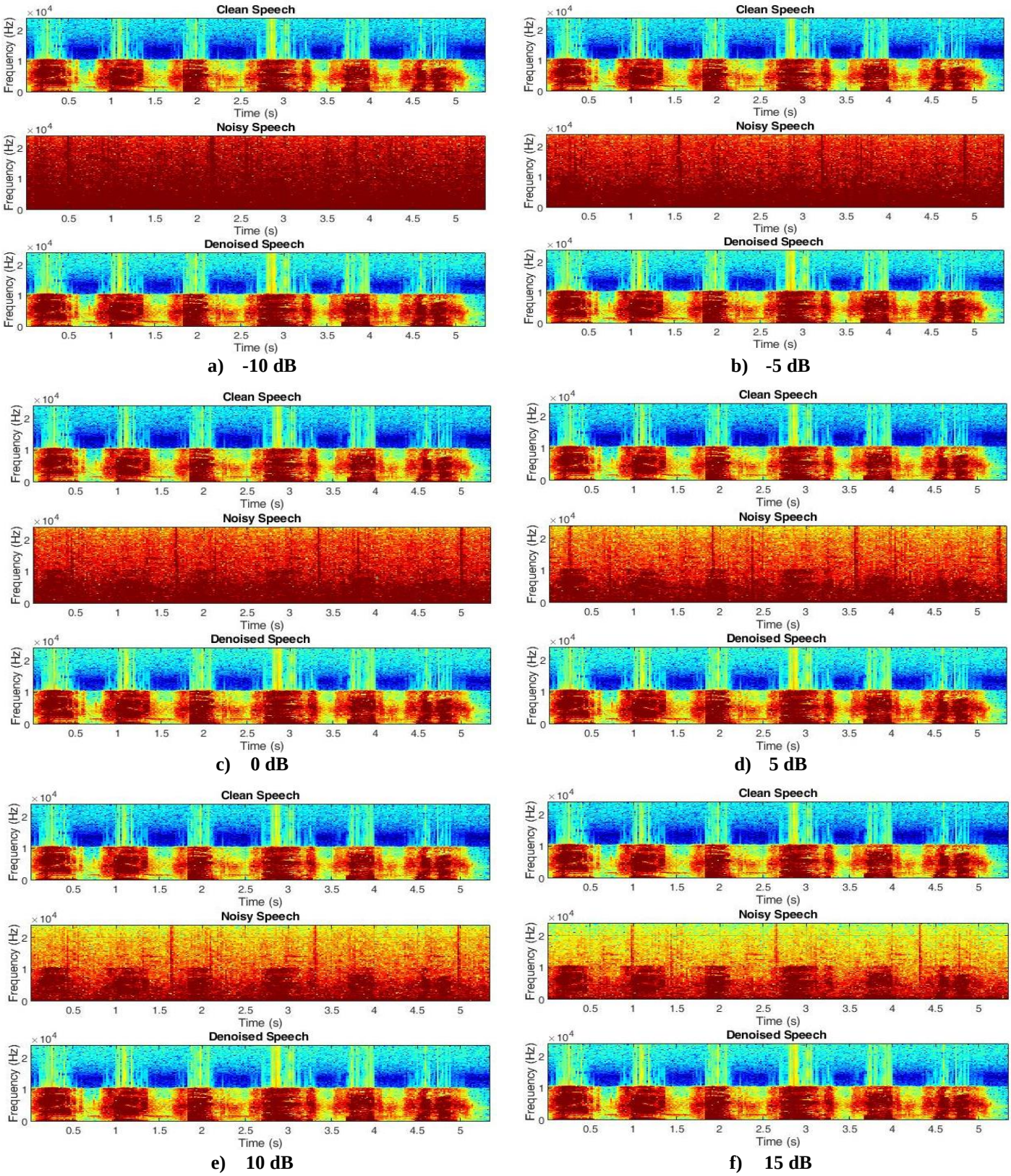


Figure 34 Modified FCRN – Spectrogram Images of Airport Noise for Alaryngeal Speech at various Noise Levels

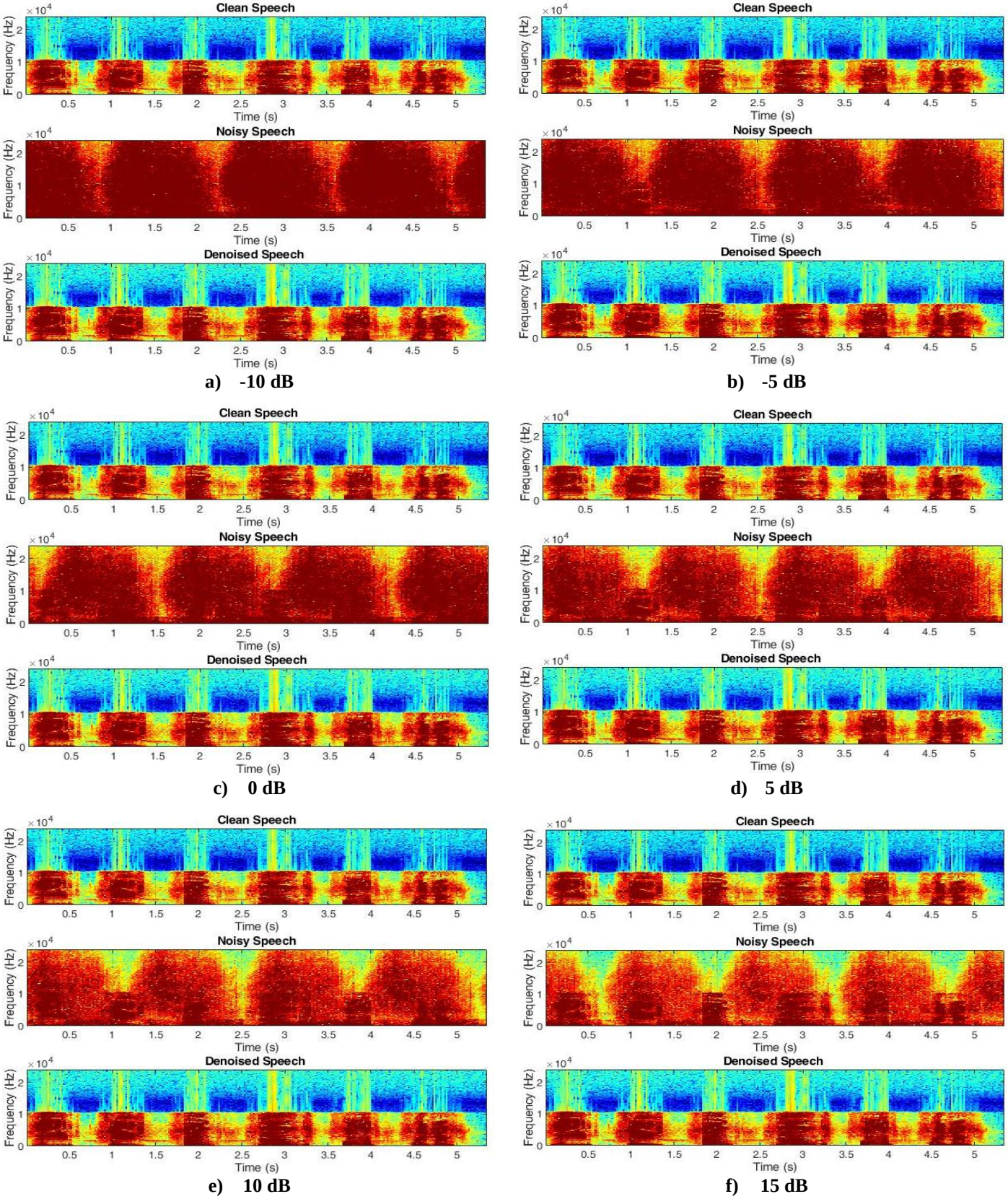


Figure 35 Modified FCRN – Spectrogram Images of Jet plane Noise for Alaryngeal Speech at various Noise Levels

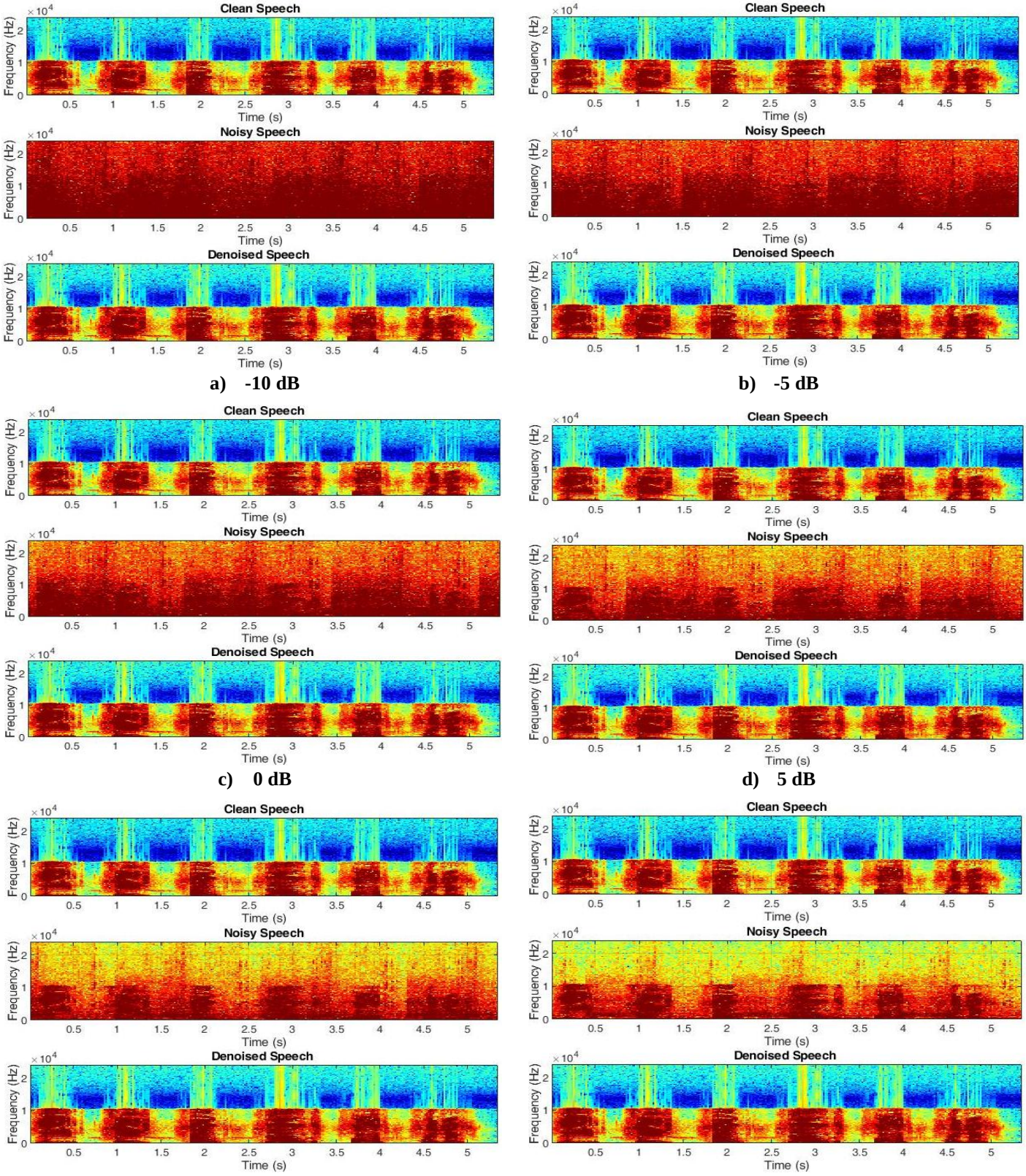


Figure 36 Modified FCRN – Spectrogram Images of Street Noise for Alaryngeal Speech at various Noise Levels

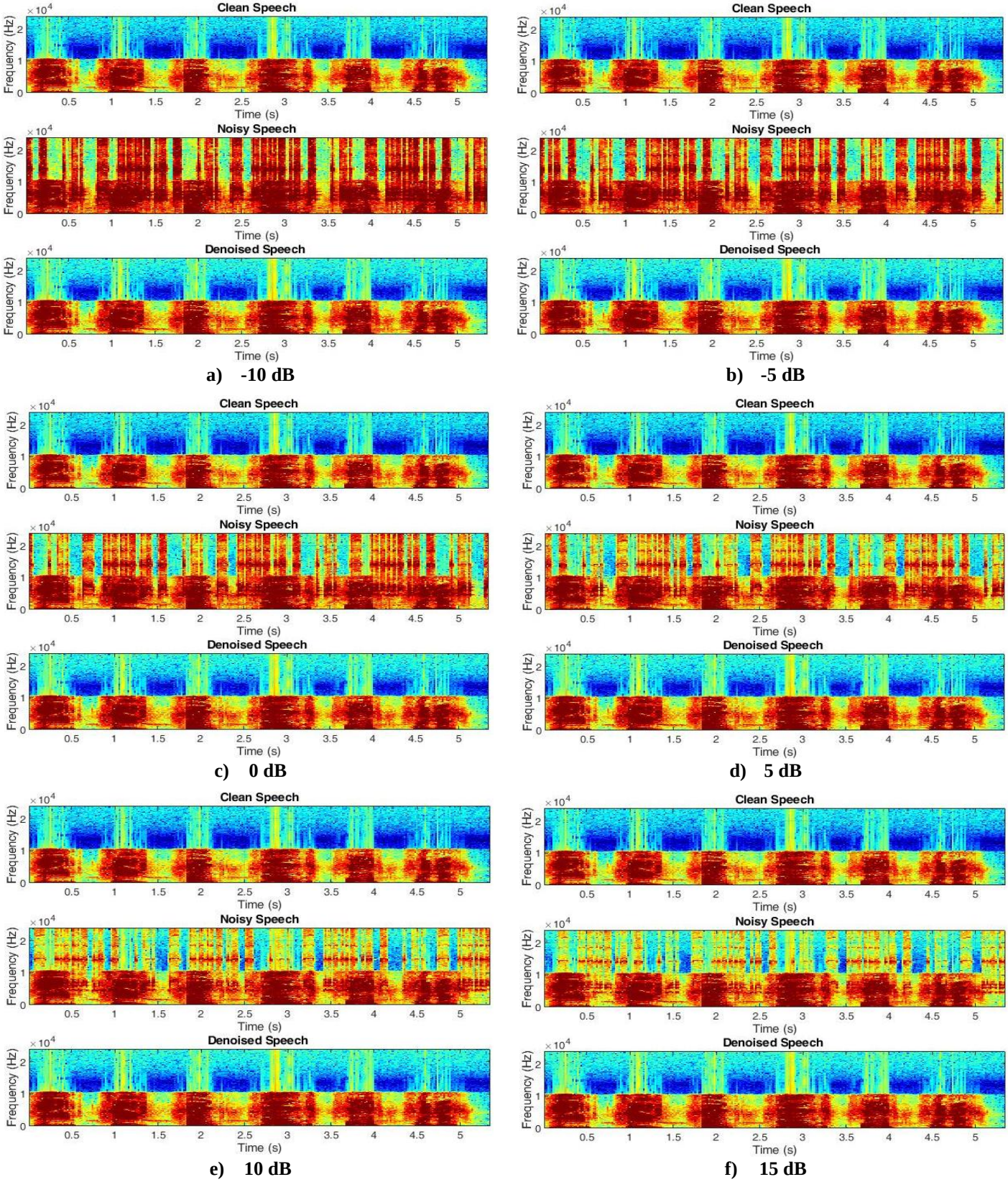
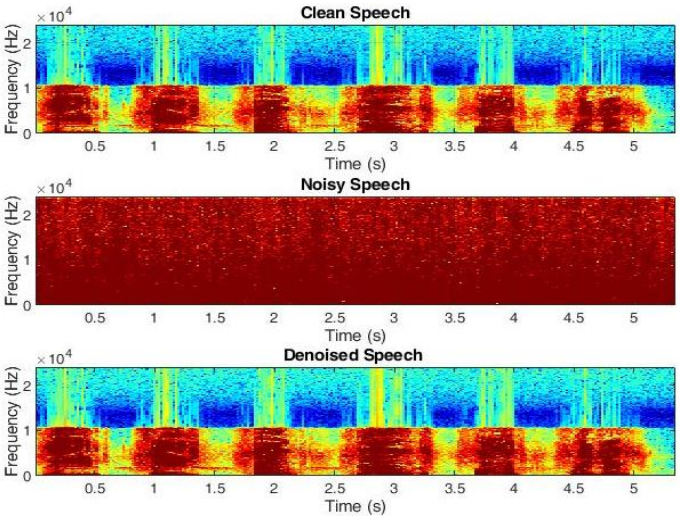
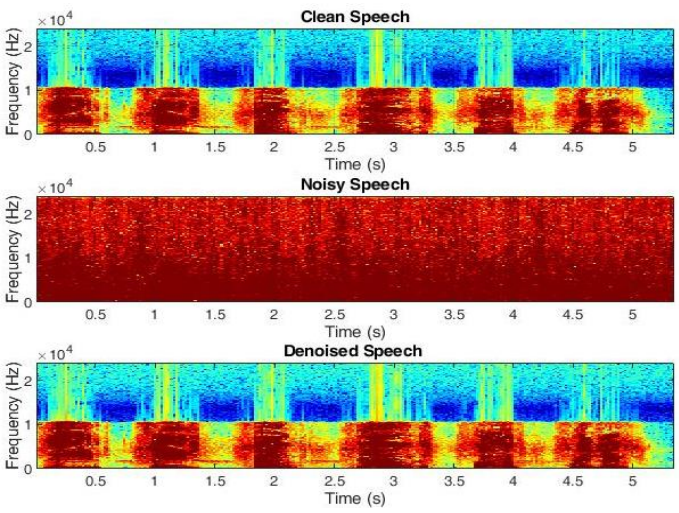


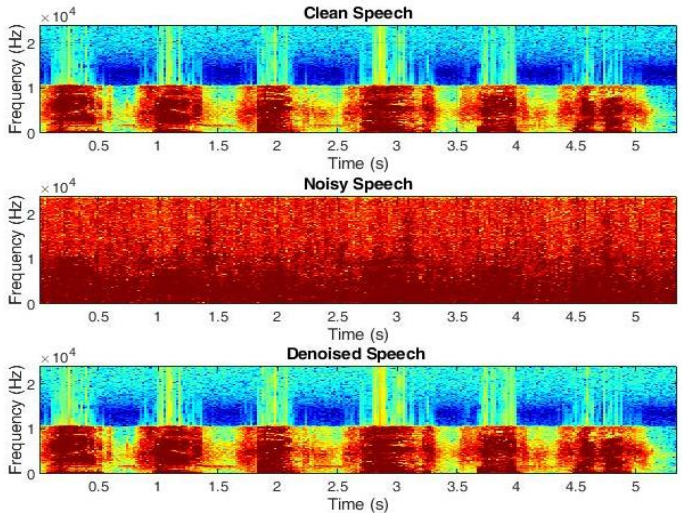
Figure 37 Modified FCRN – Spectrogram Images of Train Whistle Noise for Alaryngeal Speech at various Noise Levels



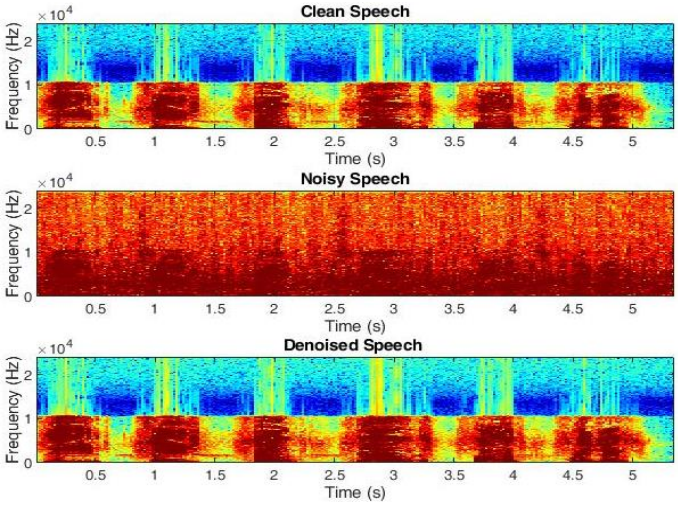
a) -10 dB



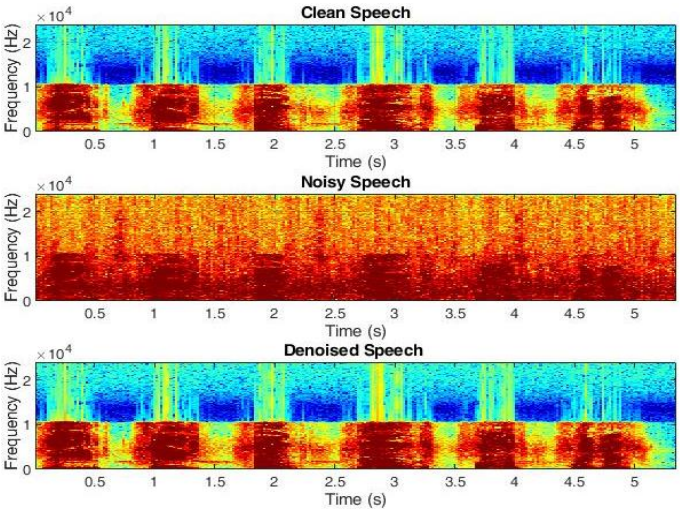
b) -5 dB



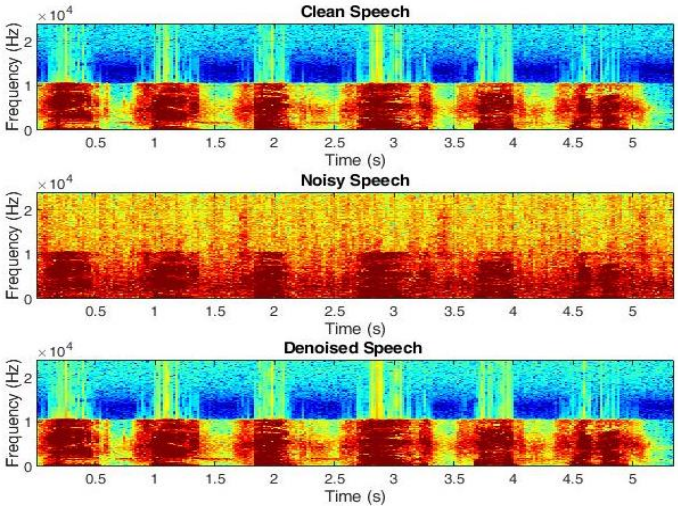
c) 0 dB



d) 5 dB

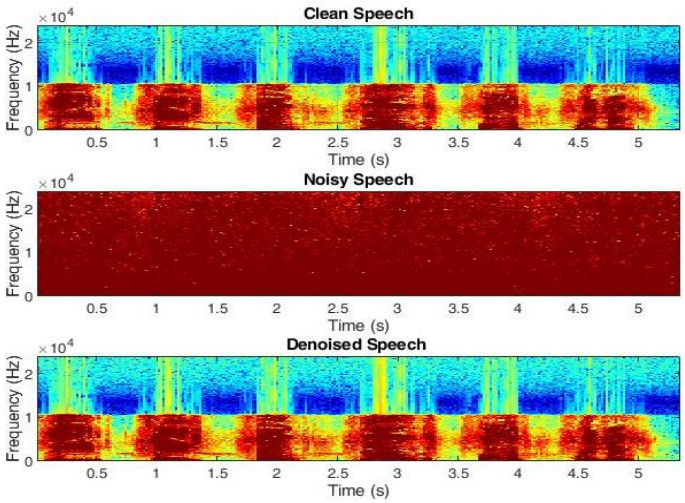


e) 10 dB

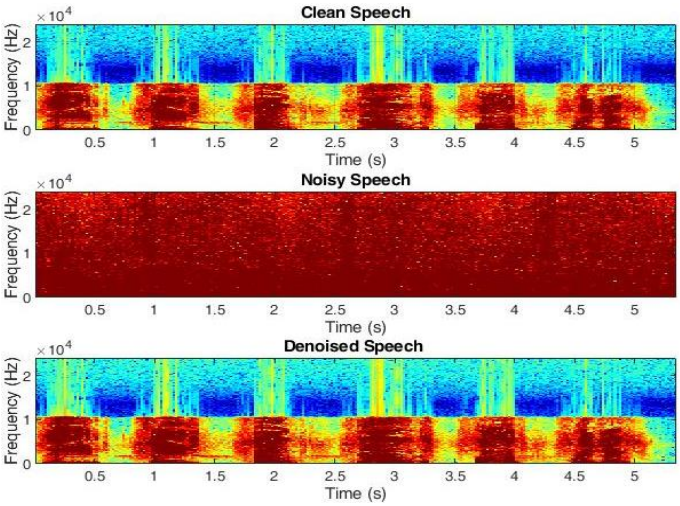


f) 15 dB

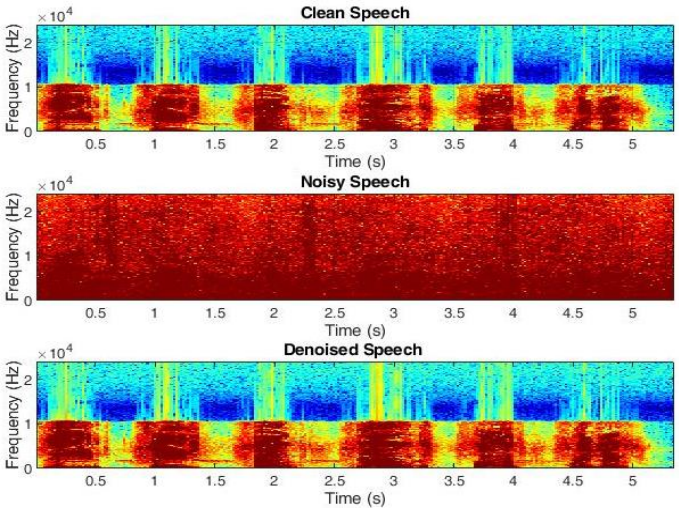
Figure 38 Modified FCRN – Spectrogram Images of Restaurant Noise for Alaryngeal Speech at various Noise Levels



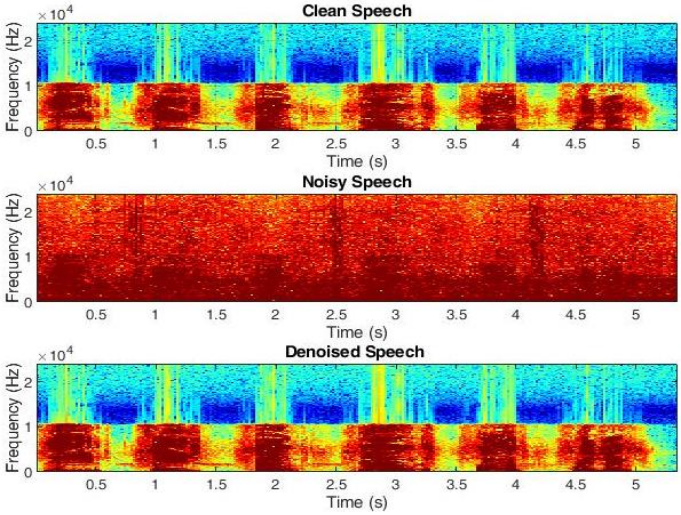
a) -10 dB



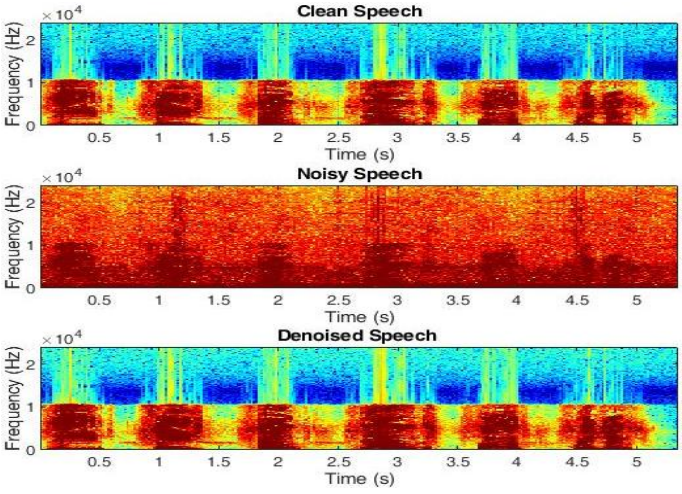
b) -5 dB



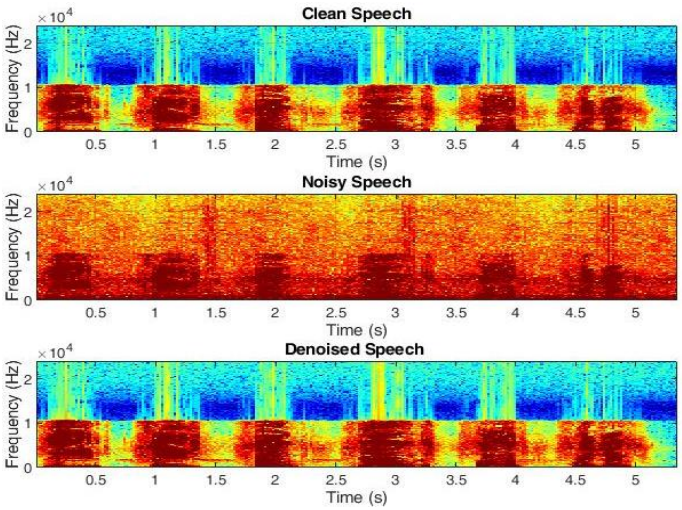
c) 0 dB



d) 5 dB



e) 10 dB



f) 15 dB

Figure 39 Modified FCRN – Spectrogram Images of Car Noise for Alaryngeal Speech at various Noise Levels

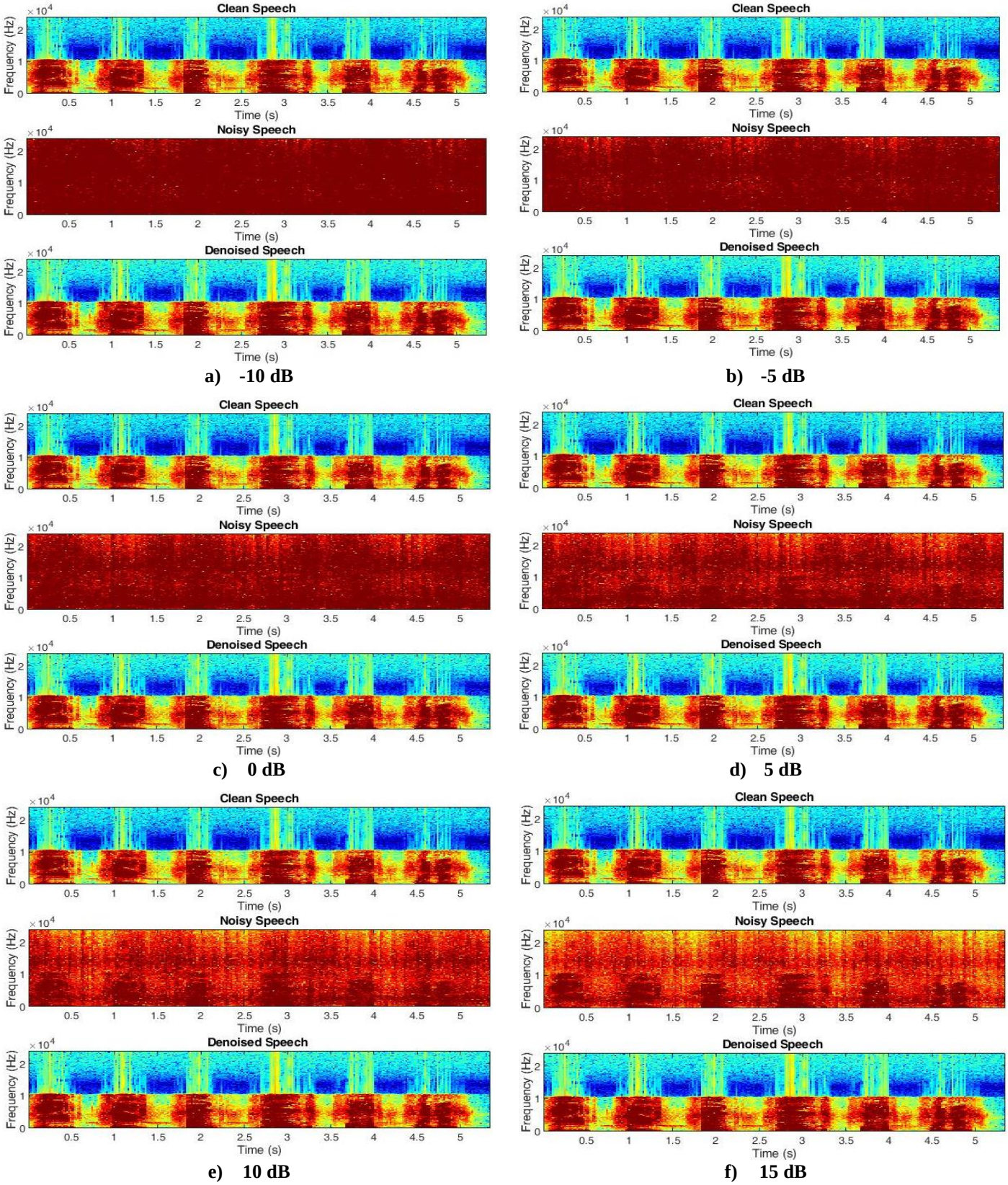


Figure 40 Modified FCRN – Spectrogram Images of Subway Noise for Alaryngeal Speech at various Noise Levels

Details of Research Publication

Name of the Scholar : HEPSIBA D

Register No : 17PHEBP001

Department : BIOMEDICAL INSTRUMENTATION ENGINEERING

Supervisor : Dr. JUDITH JUSTIN

Sl.No	Title of the Article	Journal Title	Volume/Issue/ Page No/Year	Published in UGC CARE/ SCOPUS Indexed / Web of Science List of Journals
1.	Enhancement of Single Channel Speech Quality and Intelligibility in multiple noise conditions using wiener filter and deep CNN	SOFT COMPUTING e-ISSN: 1433-7479 DOI: 10.1007/s00500-021-06291-2	Published Online 6th October 2021	SCOPUS Indexed
2.	Computational Intelligence for Speech Enhancement using Deep Neural Network	COMPUTER ASSISTED METHODS IN ENGINEERING AND SCIENCE Print ISSN: 2299-3649 DOI: 10.24423/cames.397	Published Online 16th March 2022	SCOPUS Indexed
3.	Role of Deep Neural Network in Speech Enhancement: A Review	ARTIFICIAL INTELLIGENCE Online ISBN: 978-981-13-9129-3 Print ISBN: 978-981-13-9128-6 DOI: 10.1007/978-981-13-9129-3_8	Published Online 5th July 2019	SCOPUS Indexed

Judith Justin 11/5/2022
Supervisor

Judith Justin 11/5/2022
Head of the Department

Sarg 11/5/22
Dean (School of Engineering) i/c

Soft Computing
https://doi.org/10.1007/s00500-021-06291-2

FOCUS

Enhancement of single channel speech quality and intelligibility in multiple noise conditions using wiener filter and deep CNN

D. Hepsiba[1,2] · Judith Justin[1]

Accepted: 15 September 2021

Abstract
Nowadays, deep neural network has become the prime approach for enhancing speech signals as it yields good results compared to the traditional methods. This paper describes the transformation in the enhanced speech signal by applying the deep convolutional neural network (Deep CNN), which can model nonlinear relationships and compare it with the Wiener filtering method, which is the best technique for speech enhancement among the traditional methods. Denoising is performed in the frequency domain and converted back to the time domain to analyze performance metrics such as speech quality and speech intelligibility. The speech quality is analyzed based on the signal to noise ratio (SNR) and perceptual evaluation of speech quality (PESQ). Speech intelligibility is analyzed by short-time objective intelligibility (STOI). Both the methods evaluated the denoised speech, and the analysis made on the results shows that the SNR of the conventional Wiener filtering method is much improved when compared with Deep CNN. However, the PESQ and STOI of Deep CNN-based enhanced speech outperform the Wiener filtering method. The performance metrics indicate that Deep CNN achieves better results than the conventional technique.

Keywords Deep convolutional neural network · Noisy speech · Speech enhancement · Speech quality · Intelligibility

1 Introduction

Communication through speech is one of the vibrant methodologies to express one person's internal thoughts to another and from the human to machine and vice versa. The original quality of the speech signal becomes distorted as it is delivered into the outside world. Therefore, the speech signal mixed with noise needs to be enhanced. Consequently, speech enrichment needs to enhance the quality and legibility (Wang et al. 2021) of noisy speech signals. The need of the hour in our day-to-day life is the extraction of the clear speech signal from the distorted noisy speech which are prone to background noise and reverberations.

Speech signal enhancement is a tedious process compared to other signals because of its characteristic that changes intensely with time. The algorithms used for this process need to give a spontaneous action for different practical applications. The most common speech processing techniques for denoising that are used for enhancing the speech signal are minimum mean square error method (Schwerin and Paliwal 2014) that is performed by short-time spectral magnitude estimation between the clean speech signal and enhanced speech signal, spectral subtraction method (Paliwal et al. 2010) that deals with the clean speech spectrum estimation by subtraction of noise spectrum from noisy speech spectrum. Various filtering techniques like Wiener filter (Grais and Erdogan 2013) that acts as a linear estimator for reducing the mean squared error (MSE) between the clean speech and enhanced speech signal, and Kalman filter (Dionelis and Brookes 2018) estimates the model from observing a set of the noisy speech signal. These statistical-based (Hu and Loizou 2008; Loizou 2013) unsupervised models are imperfect in

Communicated by Joy Iong-Zong Chen.

✉ D. Hepsiba
hepsiba@karunya.edu

Judith Justin
hod_bmie@avinuty.ac.in

[1] Department of Biomedical Instrumentation Engineering, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, Tamil Nadu, India

[2] Department of Biomedical Engineering, Karunya Institute of Technology and Sciences, Coimbatore, Tamil Nadu, India

Published online: 06 October 2021

predicting the variations because of dynamic nature of noisy speech signals. The statistical assumptions need to be made in the unsupervised models and that does not improve the performance of the denoised speech. The supervised models are data driven and it eliminates the statistical assumptions that are made on the clean and noisy speech signals.

Nowadays, the enhancement techniques incorporate the taxonomy of artificial intelligence by which the machine learning (Srinivasan et al. 2006) and deep learning technique (Kolbk et al. 2017; Wang and Chen 2018; Chai et al. 2019) is widely applied to improve the clarity of speech (intelligibility) and it increases the listening capability (based on quality) so that it is perceived. It is imperative as the listeners are interested and focused on listening to the speech signal with excellent quality and intelligibility. Denoising is a fundamental strategy that is implemented for applications that deal with speech signals such as telecommunication (Rix et al. 2001), speaker recognition in biometrics (Jain et al. 2004), hearing aids (Healy et al. 2017), hands-free communication (Thiergart and Taseska 2014) and many more.

The drawbacks of the unsupervised techniques could be overcome by applying the deep neural network that deals with training the network with massive data in multiple noise conditions. The data-driven approach (Zhao et al. 2018) of the deep neural network makes it more efficient and is responsive to untrained conditions and unseen noises. In the recent past, the commonly used techniques for supervised speech enhancement (Nossier et al. 2021) technique include the mapping in the frequency domain or time–frequency masking. The speech signal is converted from the frequency domain to the time domain. These methodologies enable the reconstruction of the speech signal from frequency domain to time domain with the phase of the noisy signal (Li et al. 2019).

The order of the content of this research paper is as follows: the recent work carried out in speech enhancement is discussed in the 2nd Section. A clear explanation of the proposed Deep CNN system and a comparison with the Wiener filter is given in the 3rd Section. Section 4 discusses the dataset used, features extracted, algorithm and its description. The description of the results obtained and the conclusion are mentioned in the 5th and 6th Section, respectively.

2 Related works

Similar works carried out in the speech enhancement area helps in removing the background noise that affects the speech signal are the weighted noise encoder for enhancing the speech signal by considering the power spectrum of clean speech and the SNR to build the Wiener filter in the frequency domain (Xia and Bao 2014). Modeling of the time and frequency correlation dimensions by applying the improved minima controlled recursive averaging (IMCRA) and also incorporating the long short-term memory (LSTM) of recurrent neural network (RNN) architecture and CNN exhibits good results in terms of the performance metrics (Yuan 2020). Cycle consistent training (Meng et al. 2018) for enhancement optimizes clean to noisy and noisy to clean speech mapping simultaneously.

The different DNN-based speech enhancement methodologies adopted vary based on neural network architecture, training the target and selection of training features. Nowadays, the deep learning models that are becoming popular in the field of speech enhancement are the CNN (Zheng et al. 2020; Li et al. 2020), LSTM (Li et al. 2019), and RNN (Xian et al. 2021), which incorporate the transformation function to convert the spectral features of the noisy speech signal and clean speech signal. As CNN is widely used for image processing and recognition, it would be a good solution for the problems faced with the degradation of speech signals due to background noise. The SNR-aware (Fu et al. 2016) CNN for the enhancement process shows that the CNN suits well for extracting the time–frequency features and moves forward in achieving the goal. Loss functions based (Fu et al. 2018; Li et al. 2020) on the performance metric STOI are used for modeling the utterance as a whole.

CNN implemented to perform end-to-end speech enhancement (Du et al. 2017) task can estimate the phase of clean speech that improves the quality and intelligibility of speech. Some of the speech enhancement methods perform direct enhancement on the raw speech waveforms by mapping (Fu et al. 2017; Pandey and Wang 2019) and are referred to as the waveform-based approaches. The fully convolutional neural network (Park and Lee 2017) is one among them that allows direct mapping and feature selection from the convolutional encoder-decoder model (Lan et al. 2020). Obtaining the mean absolute error loss for the training of CNN is done by taking the magnitude of the enhanced STFT and clean STFT (Pandey and Wang 2019). In some cases, a combination of the CNN and RNN model (Hsieh et al. 2020) works out to be more suitable to capture the local and sequential correlations (Wang et al. 2021). Another approach uses sequence to sequence model (Kameoka et al. 2020) using LSTM RNN to model the encoder by encoding the input sequence and decoder to decode the output sequence for voice conversion.

The mapping function created based on the noisy and clean speech signal by the nonlinear-based regression model (Xu et al. 2013) shows that the ability to handle the unseen noise is diminished. In the ILMSAF-based speech enhancement, the performance of the network is reduced

for the volvo noise (Li et al. 2016; Sungheetha and Rajesh 2021; Kumar 2021). As the task is to enhance the speech signal by removing the noise, the CNN is applied for the speech enhancement as it was observed that it gives improved results compared to multi-layer perceptron (Grais and Plumbley 2017).

The CNN is robust and suits well for speech enhancement. Therefore, in the proposed work, the Deep CNN is designed to give outperforming results. Deep CNN takes the noisy speech signal as the input and converts it into the frequency domain to train the network. It is because the noise and the clean speech signal can be discriminated only in the frequency domain. The training is performed until the mean squared error is minimum between the clean speech signal and the denoised or enhanced speech signal.

3 Speech enhancement system

In today's scenario, the best of all techniques are the Deep algorithms, as they can handle a lot of data and design a model by themselves. In this work, the Deep CNN is designed to perform speech enhancement and a comparative study is done by analyzing its performance with the best conventional technique, i.e., the Wiener filter as shown in Fig. 1. Therefore, the best conventional Wiener filter and Deep CNN are taken for comparison. The comparison results show that each technique is best in its way.

3.1 Model of speech signal

The noisy speech signal is acquired from adding the clean speech signal with the different types of noise as given in Eq. 1. The task is to retrieve the clean speech signal from the noisy speech signal by eliminating the noise.

$c(n)$: Clean Speech Signal
$b(n)$: Noise Signal
$s(n)$: Noisy Speech Signal

$$s(n) = c(n) + b(n) \tag{1}$$

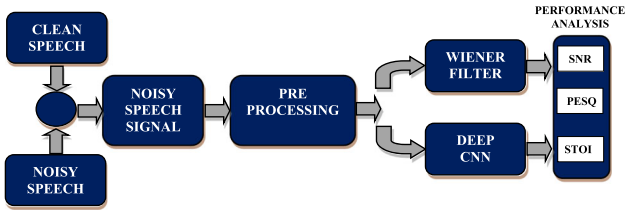


Fig. 1 Speech enhancement system for performance analysis

3.2 Wiener filtering

The presence of noise is unavoidable in real-world scenarios of speech processing. The most fundamental methodology in noise reduction of a speech signal is the optimal Wiener filter. The Wiener filter acts as a linear filter that could be utilized to separate the clean speech signal from the noisy speech signal by reducing the MSE between the estimated signal and the original signal. As the Wiener filter can achieve noise reduction, it also has the disadvantage of losing the speech signal's integrity. Therefore, the speech misrepresentation should be managed in such a way by adequately manipulating the Wiener filter or to have explicit knowledge of the speech signal. In any speech communication system, the speech signal could be distorted by background noise and reverberations. Therefore, noise reduction methodologies and speech enhancing techniques are needed to obtain the desired speech signal from the corrupted ones.

$$R(\omega) = \frac{C(\omega)}{S(\omega)} = \frac{C(\omega)}{C(\omega) + B(\omega)} \tag{2}$$

where $C(\omega)$—Signal Spectrum, $B(\omega)$—Noise Power Spectrum, $S(\omega)$—Noisy Speech Spectrum

$$R_{\text{Wiener}}(\omega) = \frac{C(\omega)}{S(\omega)} = \frac{S(\omega) - B(\omega)}{S(\omega)} \tag{3}$$

E_s—Estimation of enhanced signal

$$\widehat{E_s}(\omega, k) = R_{\text{Wiener}}(\omega) S(\omega, k) \tag{4}$$

$$\left|\hat{d}[n]\right| = \text{IFFT}\left[\sqrt{\widehat{E_s}(\omega, k)}\right] \tag{5}$$

By combining the magnitude of the clear speech spectral data with the phase of the noisy speech, the estimate of the enhanced speech is obtained. It is given as,

$$\left|\hat{d}[n]\right| = \left|\hat{d}[n]\right| \angle s[n] \tag{6}$$

$\hat{d}$—Estimate of Enhanced Speech.

3.3 Speech denoising and enhancement using deep convolutional neural network

Deep learning adopts the learning methodologies to create a model based on the data given to it. Neural network is the basic building block of deep learning. Speech enhancement is much required as the speech signal gets easily corrupted due to multiple noise conditions and noise levels. The noises can be stationary or nonstationary with varying acoustic characteristics. As the DNN can possess the model of highly nonlinear parameters, it makes the speech enhancement process simpler. The DNN architecture adopts the multi-layer feedforward network. The Deep

CNN is designed with multiple hidden layers with rectified linear unit (ReLU) activation function for speech enhancement. The input applied to the Deep CNN system is the frames of the noisy speech signal, and the expected output is the denoised speech signal.

The clean and noisy speech signal is converted to the frequency domain using STFT. The magnitude spectrum of the clean speech signal is taken as the target. The noisy speech signal is taken as the predictor and presented to the Deep CNN for denoising the speech as shown in Fig. 2. The regression network uses the magnitude of the noisy speech signal to reduce the mean square error between the denoised speech signal and the clean speech signal. The output from the Deep CNN gives the denoised signal in the frequency domain. The denoised speech signal is converted to the time domain using the output magnitude spectrum from the Deep CNN network and the phase of the noisy speech signal.

4 Algorithm description

Algorithm

- **Adding noises of various levels 0dB, 5dB, 10dB, 15dB to the clean speech**
- **Apply STFT to generate magnitude STFT vectors from clean and noisy speech signals**

$$X(w,n) = \sum_{m=-\infty}^{\infty} x(m)w(n-m)e^{-jwm}$$

- **Extract the Magnitude STFT feature of clean and noisy speech**
- **Normalize the feature to zero mean and unity standard deviation**
- **Holdout validation for splitting training and testing data**
- **Apply Deep Convolution Neural Network**
- **Compare performance metrics SNR, PESQ and STOI of denoised speech with noisy speech**

4.1 Dataset

The clean speech signal is taken from the University of Edinburgh, Centre for Speech Technology Research (CSTR) (https://datashare.is.ed.ac.uk/handle/10283/2791). The dataset contains nearly 400 speech sentences. These speech sentences are taken for training and different types of noise are added with different decibels. The dataset is divided into training and testing data by applying holdout validation method. 80% of the dataset is taken as training data and 20% is taken as the testing data. The noisy speech signal is generated for feeding the Deep CNN. The noisy data set is created by mixing the clean speech with the different noise types such as washing machine noise, rainbow noise, jet airplane noise and train whistle noise with different noise levels such as 0 dB, 5 dB, 10 dB and 15 dB.

The dataset contains 400 utterances and it is split into 3:1 for training and testing. Deep CNN is trained with 300 sentences and tested with 100 sentences. The training set is created by mixing the noise with the clean speech signal at different noise levels. From the testing set, the noisy speech signal is randomly chosen to check the denoising ability of the network.

4.2 Feature extraction

The first step is to convert the speech signal from the time domain to the frequency domain using STFT to extract features. The magnitude STFT vectors of the clean speech and the noisy speech are input features to the Deep CNN Model. Therefore, the speech signal is divided into a 10 ms frame with no frameshift. In converting from the time domain to frequency domain using STFT, the hamming window is utilized with a window length of 256 samples and 75% overlap. For training and testing purposes, the speech signal is down sampled to an 8 kHz signal and a 256-point FFT is implemented and the number of frequency bins is 129. The clean speech corpus taken from the open-source dataset was contaminated by the noise signals at different noise levels.

The discrete Fourier transform is applied on the overlapped frames for acquiring the STFT of the signal. Due to the overlap, the successive frames cause the nearby frames to have common samples at the boundary of the overlap.

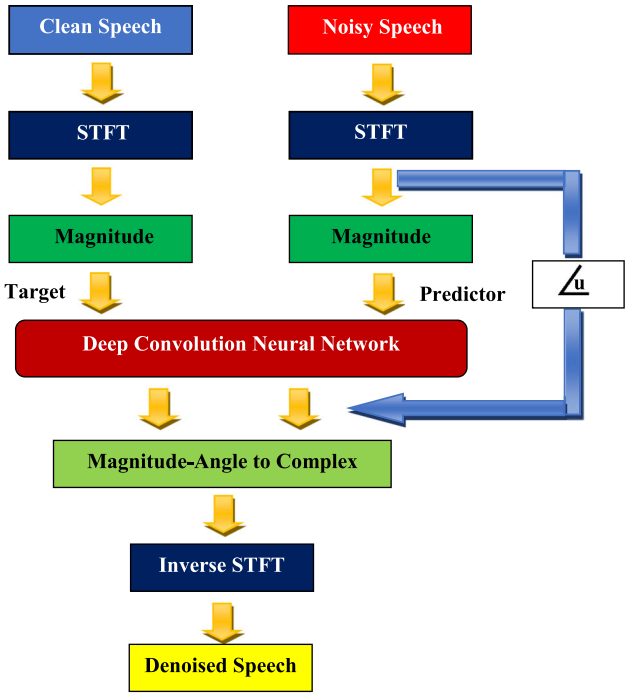


Fig. 2 Proposed deep CNN for speech enhancement

The relationship between STFT magnitude and the STFT phase is due to the correlation between the adjacent frames in the frequency domain. The original speech signal is reconstructed by maintaining a relation between the STFT magnitude and phase.

4.3 Denoising using convolutional layers

The denoising algorithm utilizes convolutional layers in which each neuron is connected to all the activations in the previous layer. Deep CNN is used to learn the spectral mapping from the noisy speech signal to the clean speech signal. The Deep CNN in this work is designed with 2-D convolutional layer and applies the sliding filter to the input as shown in Fig. 3.

The inputs to the convolutional layer are the features taken from the magnitude vector of STFT and the number of segments of the noisy speech signal. The convolution layer convolves by moving the filter on the input vertically and horizontally. The dot product is determined by the weights and the input and it is added to the bias.

The convolutional layers are defined as a group of layers, i.e., Convolutional Layer, Batch Normalization Layer and ReLu Layer and repeated 6 times, with the filter width of 9, 5 and 9 and the number of filters are 18, 30 and 8. The final convolutional layer is given a filter width of 129 along with 1 filter. The mean and standard deviation of outputs are normalized using the Batch Normalization Layers. The maximum epoch is set to 15; therefore, the network makes 15 passes through the training data. The shuffle is made for the training sequence at the starting of every epoch. During the training phase, the Adam Optimizer is used for optimizing the parameters and the MSE is taken as the loss function.

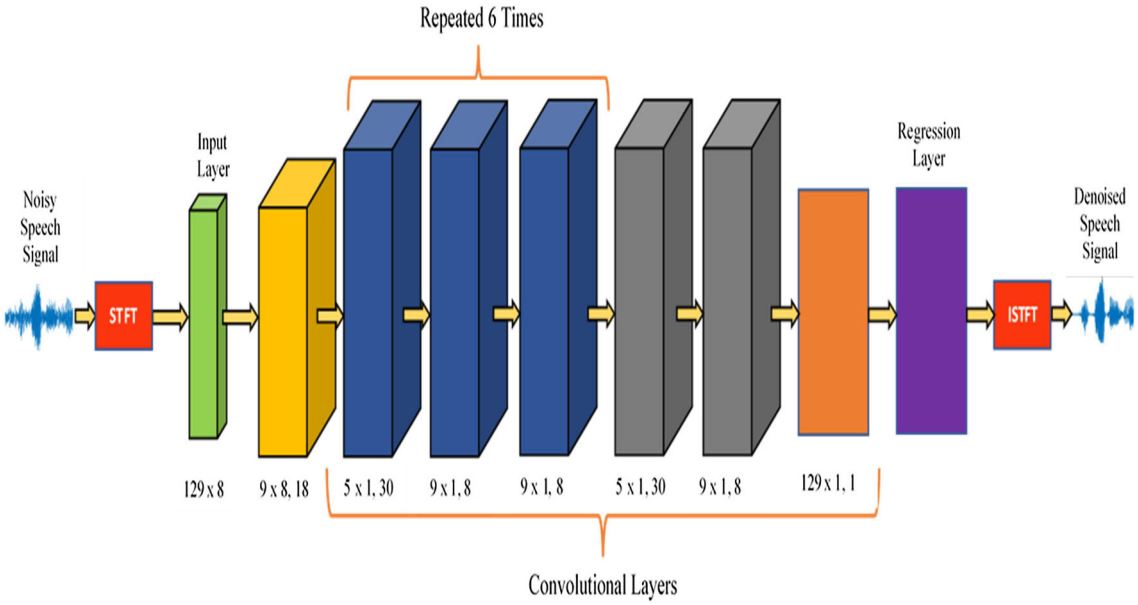


Fig. 3 Deep CNN architecture for denoising speech signal

Table 1 PESQ description

PESQ Score	Description
4–5	Excellent
3–4	Good
2–3	Fair
1–2	Poor
0–1	Bad

5 Results and discussions

The clean speech signal is added with different noise types such as washing machine noise, rainbow noise, train whistle noise and jet airplane noise with different noise levels such as 0 dB, 5 dB, 10 dB and 15 dB. The noisy speech signal generated by adding washing machine noise is given as input to the Wiener filter and DNN-based speech enhancement system. The SNR of the denoised signal is improved compared to the SNR of the noisy signal.

Fig. 4 Performance improvement comparison for noise levels 0 dB, 5 dB, 10 dB, 15 dB **a** washing machine noise, **b** rainbow noise, **c** train whistle noise, **d** jet airplane noise

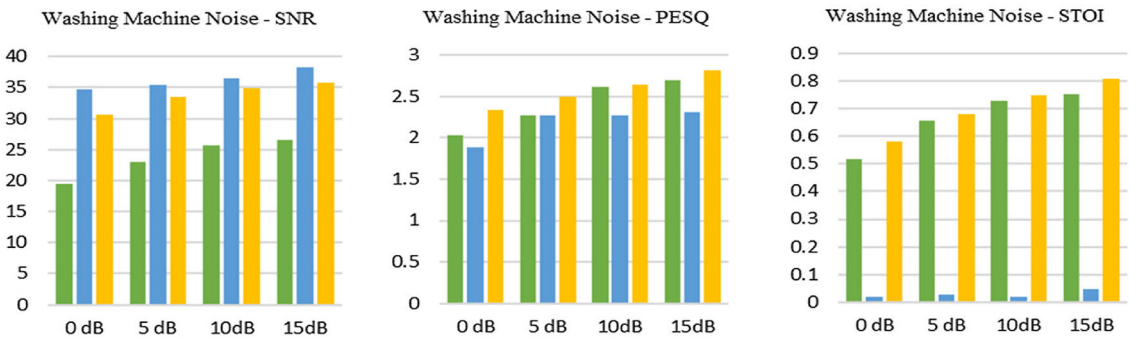
The noisy signals are taken for different noise levels such as 0 dB, 5 dB, 10 dB and 15 dB for the different noise types and were added with the clean speech signal to form the noisy speech signal. For analyzing the enhanced speech signal, the performance metrics considered are SNR, PESQ and STOI.

The performance metrics are calculated as follows:

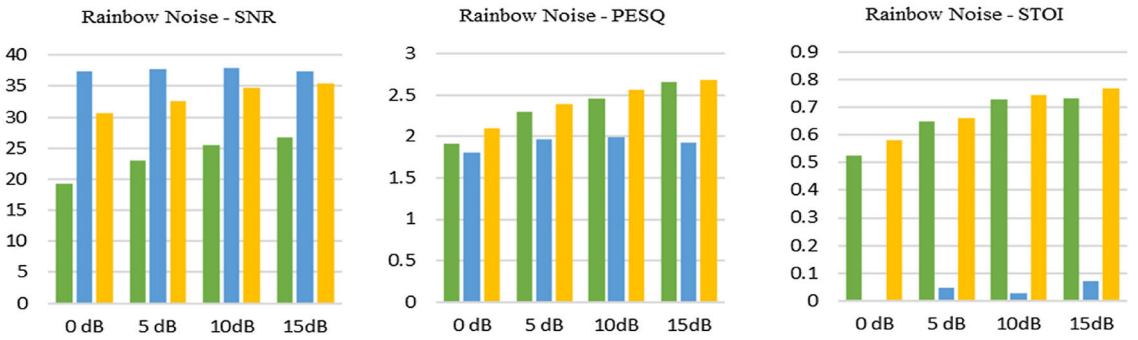
- Signal to Noise Ratio (SNR)

Table 2 Comparison of SNR, PESQ and STOI of noisy signal and denoised signal using Wiener filter and deep CNN

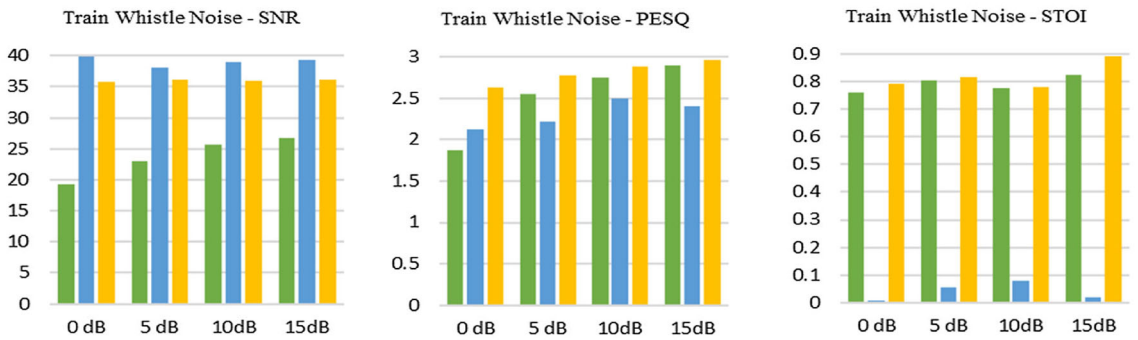
Noise level (dB)	Washing machine noise			Rainbow noise			Train whistle noise			Jet airplane noise		
	SNR			SNR			SNR			SNR		
	Noisy signal	Wiener filter	Deep CNN	Noisy signal	Wiener filter	Deep CNN	Noisy signal	Wiener filter	Deep CNN	Noisy signal	Wiener filter	Deep CNN
0	19.4898	34.7837	30.5801	19.3679	37 .2929	30.5738	19.2925	39.801	35.7682	19.5242	36.4482	33.2925
5	23.0452	35.3452	33.4311	22.9899	37.7086	32.5593	23.0805	38.0964	36.0604	23.1936	38.9859	34.2398
10	25.5831	36.5445	34.8537	25.5454	37.9403	34.6552	25.6373	38.9824	36.0091	25.5068	44.103	35.1301
15	26.6226	38.3302	35.7433	26.7152	37.4407	35.5169	26.6514	39.346	36.1191	27.0348	42.8054	36.1123
Noise level (dB)	Washing machine			Rainbow			Train whistle			Jet airplane		
	PESQ			PESQ			PESQ			PESQ		
	Noisy signal	Wiener filter	Deep CNN	Noisy signal	Wiener filter	Deep CNN	Noisy signal	Wiener filter	Deep CNN	Noisy signal	Wiener filter	Deep CNN
0	2.0374	1.8916	2.3326	1.91	1.812	2.0966	1.8663	2.1188	2.6372	2.2741	1.5465	2.3693
5	2.2703	2.2672	2.4992	2.2918	1.9707	2.3908	2.5498	2.2221	2.7768	2.5966	1.6103	2.6944
10	2.6184	2.2699	2.6496	2.4612	1.995	2.56	2.7473	2.4913	2.8795	2.6331	1.7719	2.8764
15	2.6983	2.3078	2.816	2.6623	1.9265	2.6776	2.9022	2.4074	2.9699	2.6785	1.8318	2.7983
Noise level (dB)	Washing machine			Rainbow			Train whistle			Jet airplane		
	STOI			STOI			STOI			STOI		
	Noisy signal	Wiener filter	Deep CNN	Noisy signal	Wiener filter	Deep CNN	Noisy signal	Wiener filter	Deep CNN	Noisy signal	wiener filter	Deep CNN
0	0.5166	0.0173	0.5809	0.5252	0.0039	0.5812	0.7598	0.0058	0.7923	0.6334	0.0718	0.6726
5	0.6569	0.0278	0.6814	0.648	0.0459	0.6609	0.8047	0.0535	0.8164	0.6951	0.0724	0.7074
10	0.7284	0.0174	0.7501	0.7291	0.0263	0.744	0.7781	0.0783	0.7814	0.7459	0.0214	0.7685
15	0.7542	0.048	0.8099	0.7331	0.0686	0.7704	0.8256	0.0177	0.8912	0.7463	0.0397	0.7693



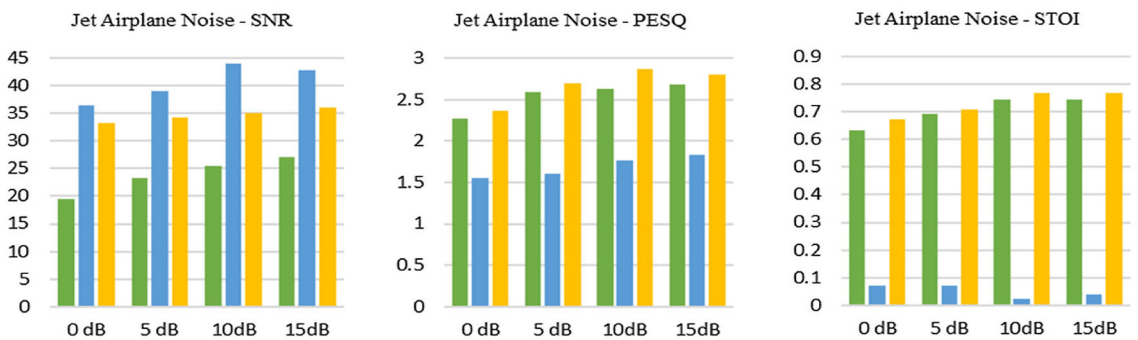
(a) Washing Machine Noise



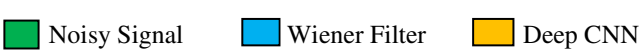
(b) Rainbow Noise

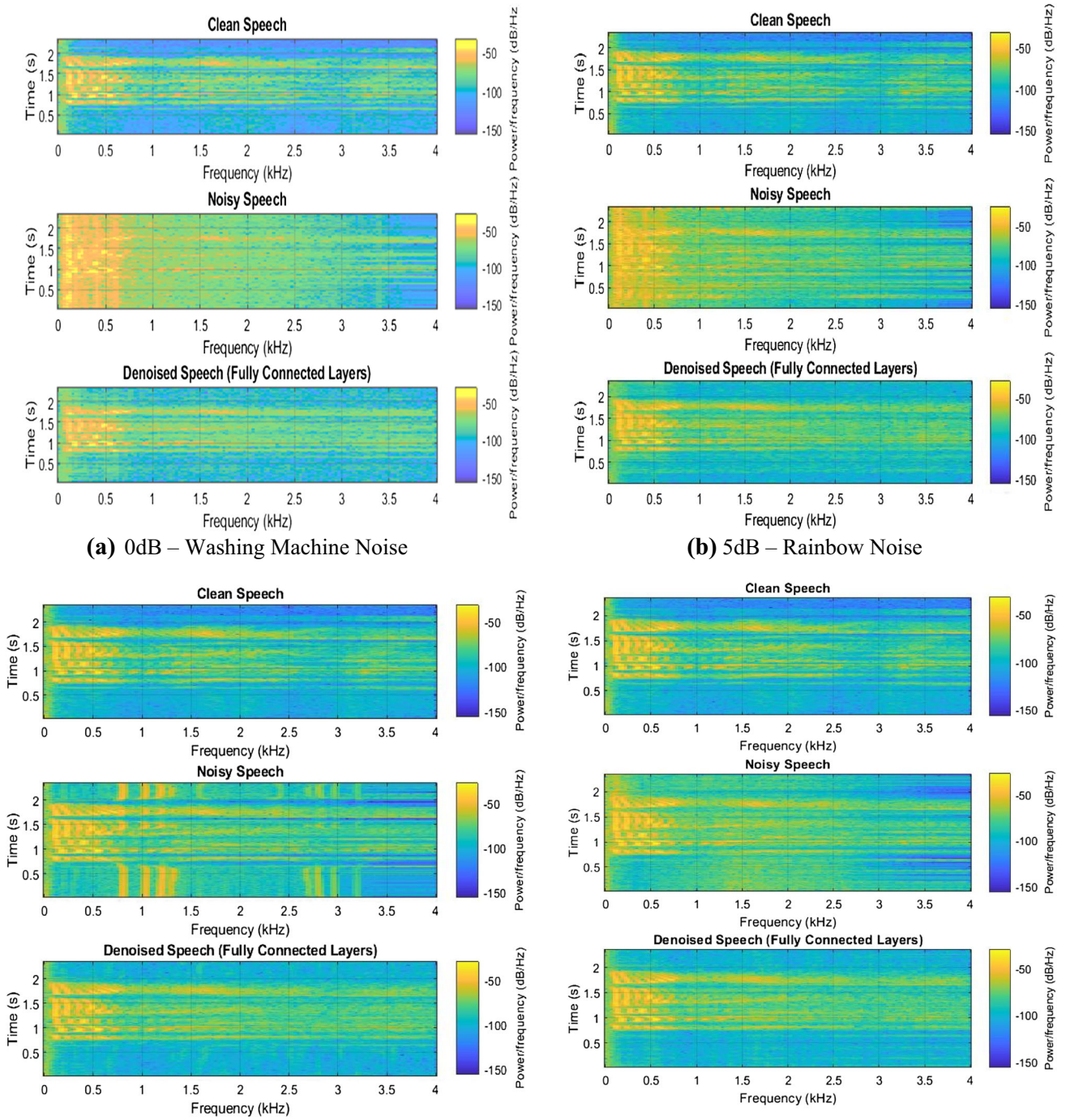


(c) Train Whistle Noise



(d) Jet Airlane Noise p





(a) 0dB – Washing Machine Noise

(b) 5dB – Rainbow Noise

(c) 10dB – Train Whistle Noise

(d) 15dB – Jet Airplane Noise

Fig. 5 Spectrogram analysis of clean speech, noisy speech and denoised speech signal **a** 0 dB—washing machine noise, **b** 5 dB—rainbow noise, **c** 10 dB—train whistle noise and **d** 15 dB—jet airplane noise

$$\mathrm{SNR}_{\mathrm{dB}} = 20\log_{10}\frac{S\_\mathrm{rms}}{N\_\mathrm{rms}}$$

where S_rms—root mean square of speech signal, N_rms—root mean square of level of noise.

- Perceptual Evaluation of Speech Quality (PESQ)

PESQ is a subjective quality measurement and it is based on the mean opinion score based on the evaluation given by the listeners and standardized by

International Telecommunications Union (ITU). The PESQ value ranges as per Table 1 given below.

- Short Time Objective Intelligibility (STOI)

STOI is a subjective intelligibility measurement, larger the value better the speech intelligibility. The STOI value ranges between 0 and 1.

The audio of the noisy speech signal was inferior in quality as well as intelligibility. When the signals were fed to the Deep CNN system for speech enhancement, the performance of the denoised speech was well improved in terms of quality which were clearly observed by the values of SNR and PESQ. Also, the intelligibility was improved, which was analyzed from the STOI scores. Table 2 shows the quality (SNR and PESQ) and intelligibility (STOI) of noisy signals and improvement in the denoised signal's performance metrics.

In order to analyze the quality, SNR and PESQ are considered and to evaluate the clarity of speech; the metric STOI is taken. The subjective quality of the spoken speech signal is analyzed by PESQ. The value of PESQ ranges between − 0.5 to 4.5. The higher the value of PESQ on the scale indicates the improvement in quality of the denoised speech. STOI refers to the subjective intelligibility of speech and it ranges between 0 and 1. The improvement in the STOI value is indicated by the higher value.

As per the observations from the performance metrics shown in Table 2, the SNR of the denoised signal through Wiener filtering shows good improvement compared to Deep CNN model for different noise levels as well as different noise types. The PESQ value of the Wiener filter is in the poor range (1–2) for the rainbow and jet airplane noise as per PESQ scores given in Table 1. But the PESQ value of the washing machine noise and train whistle noise of the Wiener filter is in the fair (2–3) range.

For the Deep CNN, the PESQ values for all the noise levels and noise types it falls in the fair (2–3) category of mean opinion score. As the Wiener filter focusses more on the quality of the speech signal, it gives good result in terms of SNR and moderate results for PESQ. But the intelligibility of speech is compromised which reduces the clarity of the speech signal. The STOI scores show that the Wiener filter is not capable of improving the intelligibility. On the other hand, the Deep CNN shows drastic results in the STOI values, which in turn represents the intelligibility of the denoised speech signal.

The consolidated results in Table 2 show the improvement in the performance metrics of Deep CNN compared to the conventional Wiener filtering algorithm for denoising speech signal. The Wiener filtering method shows outstanding results on the SNR and the PESQ. It is clearly observed that the Wiener filter has good capability in improving the quality of the speech signal. When the intelligibility of the speech signal is considered, the performance of the Wiener filter is deficient. However, the DNN shows a drastic increase in terms of the clarity of the speech signal.

The denoised signal shown in Fig. 4 represents that the SNR of the noisy signal is much improved in the Wiener filter compared to the Deep CNN. However, in terms of the other performance metric representing the quality of speech, i.e., the PESQ of the denoised speech signal is much improved in Deep CNN compared to the Wiener filter. When the intelligibility of the denoised speech is analyzed, it is evident that the STOI scores of the Deep CNN give an excellent improvement in the clarity of speech. The spectrograms of the clean speech, noisy speech and denoised speech for the different types of noise and noise levels are shown in Fig. 5.

6 Conclusion

The proposed single channel speech enhancement system estimates the magnitude of the speech signal in the frequency domain. The Deep CNN-based single channel speech enhancement system is compared with the traditional Wiener filtering method. Evaluation is carried out on multiple noise conditions to analyze the denoising capability of the speech enhancement system, and the results indicate that the Deep CNN-based system outperforms in terms of quality and intelligibility compared to the best performing Wiener filtering traditional technique. The quality of the denoised speech signal based on the SNR shows a drastic improvement for the Wiener filtered denoised signal. However, the Deep CNN yields excellent results in terms of quality and intelligibility that are analyzed based on the scores of PESQ and STOI. Thus, it should be recorded that the performance of Deep CNN outperforms the traditional Wiener filter technique.

Funding No funding.

Declarations

Conflict of interest We don't have any conflict of interest.

Human and animal rights statement Humans/animals are not involved in this research work.

Data availability statements The datasets analyzed during the current study are available in the University of Edinburgh, Centre for Speech Technology Research (CSTR). https://datashare.is.ed.ac.uk/handle/10283/2791.

References

Chai L, Du J, Liu Q-F, Lee C-H (2019) Using generalized Gaussian distributions to improve regression error modeling for deep learning-based speech enhancement. IEEE ACM Trans Audio Speech Lang Process 27(12):1919–1931

Cui X, Chen Z, Yin F (2020) Speech enhancement based on simple recurrent unit network. Appl Acoust 157:107019

De S, Smith SL (2020) Batch normalization biases deep residual networks towards shallow paths. CoRR, vol. abs/2002.10444

Dionelis N, Brookes M (2018) Phase aware single channel speech enhancement with modulation domain Kalman filtering. IEEE ACM Trans Audio Speech Lang Process 26:5

Du et al (2017) Stacked convolutional denoising auto-encoders for feature representation. IEEE Trans Cybern 47(4):1017–1027

Fu S-W, Tsao Y, Lu X (2016) Snr-aware convolutional neural network modeling for speech enhancement. In: Interspeech, pp 3768–3772

Fu S-W, Tsao Y, Lu X, Kawai H (2017) Raw waveform-based speech enhancement by fully convolutional networks. In: Proceedings of the APSIPA ASC, pp 6–12

Fu S-W, Wang T-W, Tsao Y, Lu X, Kawai H (2018) End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks. IEEE ACM Trans Audio Speech Lang Process (TASLP) 26(9):1570–1584

Grais EM, Erdogan H (2013) Discriminative nonnegative dictionary learning using cross-coherence penalties for single channel source separation. In: Proc. Inter-speech

Grais EM, Plumbley MD (2017) Single channel audio source separation using convolutional denoising autoencoders. In: Proceedings of the IEEE global conference on signal information processing, pp 1265–1269

Healy EW, Delfarah M, Vasko JL, Carter BL, Wang D (2017) An algorithm to increase intelligibility for hearing-impaired listeners in the presence of a competing talker. J Acoust Soc Am 141(6):4230–4239

Hsieh T-A, Wang H-M, Lu X, Tsao Y (2020) WaveCRN: an efficient convolutional recurrent neural network for end-to-end speech enhancement. IEEE Signal Process Lett 27:2149

https://datashare.is.ed.ac.uk/handle/10283/2791

Hu Y, Loizou PC (2008) Evaluation of objective quality measures for speech enhancement. IEEE Trans Audio Speech Lang Process 16(1):229–238

ITU, Perceptual Evaluation of Speech Quality (PESQ): an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs ITU-T Rec. p 862 (2000)

Jain K, Ross A, Prabhakar S (2004) An introduction to biometric recognition. IEEE Trans Circuits Syst Video Technol 14(1):4–20

Kameoka H, Tanaka K, Kwasny D, Kaneko T, Hojo N (2020) ConvS2S-VC: fully convolutional sequence-to-sequence voice conversion. IEEE ACM Trans Audio Speech Lang Process 28:1849–1863

Kolbæk M, Tran Z-H, Jensen SH, Jensen J (2020) On loss functions for supervised monaural time-domain speech enhancement. IEEE ACM Trans Audio Speech Lang Process 28:825–838

Kolbk M, Tan Z, Jensen J (2017) Speech intelligibility potential of general and specialized deep neural network-based speech enhancement systems. IEEE ACM Trans Audio Speech Lang Process 25(1):153–167

Kumar TS (2021) Construction of hybrid deep learning model for predicting children behavior based on their emotional reaction. J Inf Technol 3(01):29–43

Lan T, Lyu Y, Ye W, Hui G, Zenglin Xu, Liu Q (2020) Combining multi-perspective attention mechanism with convolutional networks for monoaural speech enhancement. IEEE Access 8:78979–78991

Li A, Yuan M, Zheng C, Li X (2020) Speech enhancement using progressive learning-based convolutional recurrent neural network. Appl Acoust 166:107347

Li R, Liu Y, Shi Y, Dong L, Cui W (2016) ILMSAF based speech enhancement with DNN and noise classification. Speech Commun 85:53–70

Li J, Zhang H, Zhang X, Li C (2019) Single channel speech enhancement using temporal convolutional recurrent neural networks. In: Proceedings of the APSIPA ASC, pp 896–900

Loizou PC (2013) Speech enhancement: theory and practice, 2nd edn. CRC Press, Boca Raton

Meng Z, Li J, Gong Y, Juang BH (2018) Cycle-consistent speech enhancement. In: Proceedings of the INTERSPEECH, pp 1165–1169

Nossier SA, Wall J, Moniri M, Glackin C, Cannings N (2021) An experimental analysis of deep learning architectures for supervised speech enhancement. Electronics 10(1):17

Paliwal KK, Wojcicki K, Schwerin B (2010) Single-channel speech enhancement using spectral subtraction in the short-time modulation domain. Speech Commun 52(5):450–475

Pandey D, Wang D (2019) TCNN: temporal convolutional neural network for real-time speech enhancement in the time domain. In: Proceedings of the Interspeech, pp 6975–6879

Pandey A, Wang D (2019) A new framework for CNN based speech enhancement in the time domain. IEEE ACM Trans Audio Speech Lang Process 27(7):1179

Park SR, Lee JW (2017) A fully convolutional neural network for speech enhancement. Proc Interspeech 2017:1993–1997

Rix W, Beerends JG, Hollier MP, Hekstra AP (2001) Perceptual evaluation of speech quality (PESQ)—a new method for speech quality assessment of telephone networks and codecs. In: Proceedings of the IEEE international conference on acoustics, speech, and signal processing, vol 2, pp 749–752

Schwerin B, Paliwal KK (2014) Using STFT real and imaginary parts of modulation signals for MMSE-based speech enhancement. Speech Commun 58:49–68

Srinivasan S, Samuelsson J, Kleijn WB (2006) Codebook driven short term predictor parameter estimation for speech enhancement. IEEE Trans Audio Speech Lang Process 14(1):163–176

Sungheetha A, Rajesh Sharma R (2021) Classification of remote sensing image scenes using double feature extraction hybrid deep learning approach. J Inf Technol 3(02):133–149

Tan K, Wang D (2020) Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement. IEEE ACM Trans Audio Speech Lang Process 28:380–390

Thiergart O, Taseska M, Habets EAP (2014) An informed parametric spatial filter based on instantaneous direction-of-arrival estimates. IEEE ACM Trans Audio Speech Lang Process 22:12

Wang D, Chen J (2018) Supervised speech separation based on deep learning: An overview. IEEE ACM Trans Audio Speech Lang Process 26(10):1702–1726

Wang NY-H, Wang H-LS, Wang F-W, Lu X, Wang H-M, Tsao Y (2021) Improving the intelligibility of speech for simulated electric and acoustic simulation using fully convolutional neural network. IEEE Trans Neural Syst Rehabil Eng 29:184–195

Xia B, Bao C (2014) Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification. Speech Commun 60:13–29

Xian Y, Sun Y, Wang W, Naqvi SM (2021) Convolutional fusion network for monaural speech enhancement. Neural Netw 143:97–107

Xu Y, Jun Du, Dai L-R, Lee C-H (2013) An experimental study on speech enhancement based on deep neural networks. IEEE Signal Process Lett 21(1):65–68

Yuan W (2020) A time–frequency smoothing neural network for speech enhancement. Speech Commun 124:75–84

Zhao H, Zarar S, Tashev I, Lee C (2018) Convolutional-recurrent neural networks for speech enhancement. In: International conference on *acoustics,* speech*,* and signal processing, pp 2401–2405

Zheng N, Shi Y, Rong W, Kang Y (2020) Effects of skip connections in CNN-based architectures for speech enhancement. J Signal Process Syst 92:875–884

CAMES, **29**(1–2): 71–85, 2022, doi: 10.24423/cames.397

This article belongs to the *Special Issue on Advanced Optimization Methods for Uncertainties in Intelligent Industrial Systems* edited by Dr. I.J. Jacob, Dr. B.H. bin Ahmad and Dr. Z.F. Khan

Computational Intelligence for Speech Enhancement using Deep Neural Network

Hepsiba D.[1),2)*], Judith JUSTIN[1)]

[1)] *Department of Biomedical Instrumentation Engineering*
Avinashilingam Institute for Home Science and Higher Education for Women
Coimbatore, Tamil Nadu, India, e-mail: hod_bmie@avinuty.ac.in

[2)] *Department of Biomedical Engineering*
Karunya Institute of Technology and Sciences
Coimbatore, Tamil Nadu, India
*Corresponding Author e-mail: hepsiba@karunya.edu

In real time, the speech signal received contains noise produced in the background and reverberations. These disturbances reduce the quality of speech; therefore, it is important to eliminate the noise and increase the intelligibility and quality of speech signal. Speech enhancement is the primary task in any real-time application that handles speech signals. In the proposed method, the most effective and challenging noise, i.e., babble noise, is removed, and the clean speech is recovered. The enhancement of the corrupted speech signal is done by applying a deep neural network-based denoising algorithm in which the ideal ratio mask is used to mask the noisy speech and separate the clean speech signal. In the proposed system, the speech signal corrupted by noise is enhanced. Evaluation of enhanced speech signal by performance metrics such as short time objective intelligibility and signal to noise ratio of the denoised speech show that the speech intelligibility and speech quality are improved by the proposed method.

Keywords: deep neural network, noisy speech, speech enhancement, feature extraction, speech quality, computational intelligence.

1. Introduction

Speech enhancement [21, 22] is very important for any speech signal suffering from distortions, reflections and background noise that varies from place to place. Therefore, the speech enhancement techniques are crucial and very important for improving the speech quality in applications such as speaker recognition, automatic speech recognition (ASR) [1, 2, 23, 40], speech coding [5, 6, 32] and hearing aids [3, 4, 32].

Speech enhancement algorithms [12, 34] help in reducing noise without disturbing the quality of target speech. When the speech quality and intelligibility are improved, it helps the listeners to listen to the speech without any restraints. The conventional algorithms of speech enhancement include minimum mean square error [7], spectral subtraction [8], Kalman filtering [11] and iterative Wiener filtering [10].

In the recent past, computational intelligence and machine learning have found wide applications in enhancing distorted speech and noise removal [38, 39]. The latest trend in speech enhancement uses deep learning [9, 37], which adopts the architecture of a deep neural network (DNN). A DNN is a feed-forward network capable of modeling relationships that are non-linear [36]. In order to model the DNN, the features such as relative spectral transformed perceptual linear prediction coefficients (RASTA-PLP) [14, 15], amplitude modulation spectrogram [30, 13], gammatone frequency cepstral coefficients (GFCC) [9, 16] and mel frequency cepstral coefficients (MFCC) [18] are extracted.

The training data consists of the speech signal with different noise types and signal-to-noise ratio (SNR) for the non-linear DNN-based regression model. The performance of the DNN is restricted to varying real-time noisy situations. Therefore, to improve the network's generalization ability, the changing nature of the noise is given as input to the network for training [31]. This helps to enhance the efficiency of the network in detecting unseen noise types.

In the past few years, a DNN has played a vital role in separating noise from speech and improving speech quality [19]. To enhance the noisy-reverberant speech [19], spectral mapping [17] is done using a single DNN, which removes noise and reverberation. Basically, the background noise causes disturbance to the clean speech. Here, denoising is performed for speech signal mixed with babble noise.

The content of the paper is arranged as follows: Sec. 2 gives the detailed description about the similar works carried out for speech enhancement, Sec. 3 discusses the methodology of speech enhancement. Section 4 explains the various feature extraction methodologies. Section 5 discusses a DNN for denoising and Sec. 6 presents the obtained results and their discussion; finally, the conclusion is given in the last section.

2. Related work

In the past years, the related work carried out for the speech enhancement has dealt with unsupervised techniques such as spectral subtraction, Kalman filtering, Weiner filtering and many more. The problem occurring in these techniques is that the method adopted for analyzing the noise is just an assumption. The disadvantages occurring in these techniques are eliminated by the powerful super-

vised technique such as codebook vectors and the model-based techniques where the speech signal and noise are known *a priori*. For the distortion-independent acoustic model, the non-matrix factorization (NMF) is more powerful in the process of separating the source in the recording made in a single-channel microphone in the existence of additive noise. The NMF-based technique [41] helps in estimating the speech signal and noise in the frequency domain. Segment-based approach [47] is another method to identify longer speech segments with its full-length speech sentence matching to remove fast-varying noise.

The previous research clearly indicates the improvement in speech enhancement performance when the features are extracted from the speech signal. The prediction of the log-power spectra (LPS) feature of the clean speech signal can be made using multi-objective learning. A long short-term memory (LSTM) technique [43] is a powerful tool that helps in a consistent improvement of the speech quality and intelligibility. The encoding of features [44] helps in the voice conversion process, and the different encoders are more effective. The usage of the deep recurrent neural network [42] is also very helpful in identifying the speech denoising system model in which the time-frequency masking is applied to one of the layers in the network.

Speech enhancements with deep learning are based on mapping or masking [45]. In the mapping-based enhancement, the relationship between the features of the noisy speech and the clean speech is considered. In the masking-based scenario, the relationship between the features of the noisy speech and the time-frequency mask is considered. The estimated mask is used to obtain the features of the enhanced speech signal. The different ideal masks for speech enhancement are ideal binary masks, ideal ratio masks and complex ideal ratio masks (cIRMs). The studies indicate that the ideal ratio mask leads to better results compared to the ideal binary mask. The cIRM [46] takes both the real and imaginary components for estimating the target.

Due to the non-linear relationship between the input and the target of the speech signal features, the networks with multiple layers and non-linear activation functions are more effective than shallow networks for the enhancement of speech signal. In certain applications, when the speech signal needs to be masked, babble noise is utilized for security purposes. In this paper, the ideal ratio mask is incorporated to obtain the enhanced speech features for denoising the speech signal affected by babble noise.

3. Speech enhancement methodology

The clean speech is mixed with the babble noise to form the noisy speech signal, and the features are extracted and given to the DNN, as shown in Fig. 1.

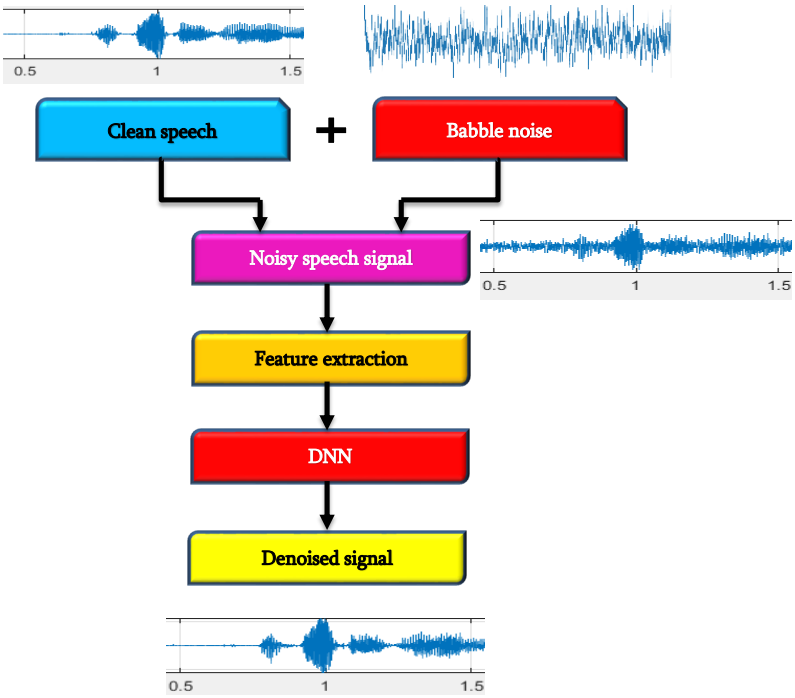


FIG. 1. Proposed speech enhancement system using a DNN.

3.1. Model of the speech signal

Let $c(t)$ and $b(t)$ represent the clean speech and babble noise, respectively. The noisy speech signal $n(t)$ is

$$n(t) = c(t) + b(t). \tag{1}$$

The babble noise signal $b(t)$ is usually not correlated with the desired signal $c(t)$; therefore, it is apparent that the noise can be removed first before recovering the clean speech. The target signal is the clean speech signal.

3.2. Process of denoising

The noisy utterance is given to the speech enhancement system, and the target signal is the noise-free clean speech. The noise is suppressed by using the time-frequency masking framework and removed by applying the time-frequency mask to the noisy speech signal. For the time-frequency masking, the ideal ratio mask is incorporated to remove the noise.

The ideal ratio mask is given by:

$$\mathrm{IRM}(t,f) = \left(\frac{C^2(t,f)}{C^2(t,f)+N^2(t,f)}\right)^{\beta}, \tag{2}$$

$$\mathrm{IRM}(t,f) = \left(\frac{\mathrm{SNR}(t,f)}{\mathrm{SNR}(t,f)+1}\right)^{\beta}, \tag{3}$$

where $C^2(t,\ f)$ shows the speech signal and $N^2(t,\ f)$ shows the noise signal, as a time-frequency (T-F) representation, and β acts as the tuning parameter for scaling the mask. At $\beta = 0.7$, the noisy signal is estimated and implemented using a DNN. After denoising, the signal is reconstructed in the time domain.

4. Feature extraction

The features extracted from the noisy speech signal are given below.

4.1. Mel frequency cepstral coefficients (MFCC)

The MFCC is the commonly used method in the feature extraction of speech signals. The speech signal is segmented into small duration blocks (windowed frames) and the fast Fourier transform (FFT) is applied to each frame sequence.

The signal is changed from the time domain signal into the frequency domain. The mel filter bank is applied to the power spectrum and energy is summed for all filter banks. The log filter bank energies are applied with a discrete cosine transform (DCT) [18, 29] to obtain the MFCC.

4.2. Relative spectral transformed perceptual linear prediction coefficients (RASTA-PLP)

RASTA-PLP is a special methodology that implements band-pass filtering to the energy in each frequency sub-band. The high-pass filter portion in the band-pass filter reduces the convolutional noise [33]. The frame-to-frame spectral changes are smoothened by the low pass portion [15].

4.3. Amplitude modulation spectrogram (AMS)

The speech signal is converted to the frequency domain by applying a short-time Fourier transform (STFT). After decomposing the signal by the bark scale decomposition, the spectral analysis is made by a second STFT. Thus, the amplitude modulation coefficients such as acoustic frequencies, time and modulation frequencies are obtained [25].

4.4. Gammatone filter bank power spectra

The input speech signal is passed through a 64 channel gammatone filter bank [24]. In each channel, the filter response is fully rectified and decimated, which is similar to windowing. The absolute values taken specify the T-F representation. The cube root of the T-F representation is taken and the DCT is applied to the cepstral coefficients [16].

4.5. Autoregressive moving average model (ARMA)

The input speech signal is taken as long segments and converted using the DCT. The windowing function is applied to the DCT signal. The ARMA modeling is applied to sub-band DCT components of the sub-band envelope.

The power spectrum estimate is yielded by integrating the sub-band envelope with respect to time. The inverse fast Fourier transform (IFFT) is used to transform the power spectrum estimates into temporal autocorrelation estimates and further used based on linear prediction in the time domain. The output obtained gives a spectrally smoothed ARMA spectrogram [19].

5. DNN for denoising

The architecture of a DNN is a feed-forward neural network [9] and has the competence to map the features of the noisy speech signal to clean the speech signal [33]. The DNN model is trained with the features extracted [9]. Figure 2 shows the DNN architecture of the proposed model.

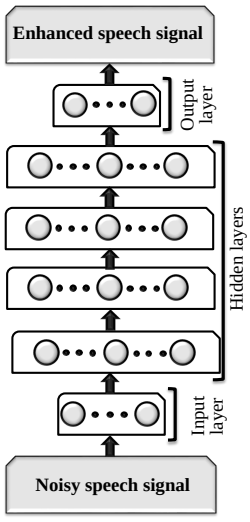


Fig. 2. The DNN architecture of the proposed model.

The sentence list is taken from the IEEE sentence database [26]. The clean and noisy audio file is taken from the Noizeus website [27]. A babble noise at 0 dB noise level is considered for mixing with the clean speech. The features extracted are 31-dimensional MFCC, 15-dimensional AMS, 64-dimensional gammatone filter bank power spectra, and 13-dimensional RASTA-PLP [35] and are taken as inputs to the DNN [19]. The DNN uses multilayer perceptron (MLP) as the discriminative learning machine, which shows good performance for speech separation. The DNN uses 4 hidden layers, each layer having 1024 rectified linear hidden units (ReLU). The number of hidden layers is taken as 4 in the process of tuning the hyperparameters as it reduces the MSE to 0.001. The network is trained with the back-propagation algorithm and the dropout rate considered is 0.2.

For the first 5 epochs, the momentum value is taken as 0.5, and after 5 epochs it is taken as 0.9. The increase of momentum rate from 0.5 to 0.9 does not fasten the training of the model, but it helps to increase the accuracy in training and testing the model. The DNN predicts the output for varying frequency ranges, and the cost function adopted is the mean squared error (MSE) [9].

For the targets in the range [0,1], the output layer uses a sigmoid activation function, and for the other layers, a linear activation function is employed. The input data given to the DNN are the features obtained from the 5-frame window for implementing the temporal context. The final estimate is obtained by finding the average of the multiple estimates of each frame [20].

6. Results and discussion

The proposed system uses sentences from the IEEE sentence database. Audio files are taken from the Noizeus website for the clean speech. The noise used for this work is a babble noise, which is the most challenging and it is considered to be the best noise for masking speech. The babble speech is generally the voice heard in the midst of the crowded ambience. The mixtures are obtained by mixing clean speech signal with babble noise with different SNR values.

The training data contains 600 sentences and the testing data consists of 120 sentences. The signal is sampled at 16 kHz and converted into frames using a 20 ms Hamming window with a 10 ms window shift for framing. For each frame, 320 frame FFT is applied, resulting in 161 frequency bins.

The SNR of the noisy speech signal shows that the noise power is greater than the signal power. After applying the denoising algorithm, the SNR is improved, which shows that the signal power has increased more than the noise power, as shown in Table 1. The noisy speech signal in the time domain, its periodogram and spectrogram are shown in Fig. 3. The spectrogram shows the intensity of noise present in the noisy speech signal. The periodogram displays the spectral density of the noisy speech signal.

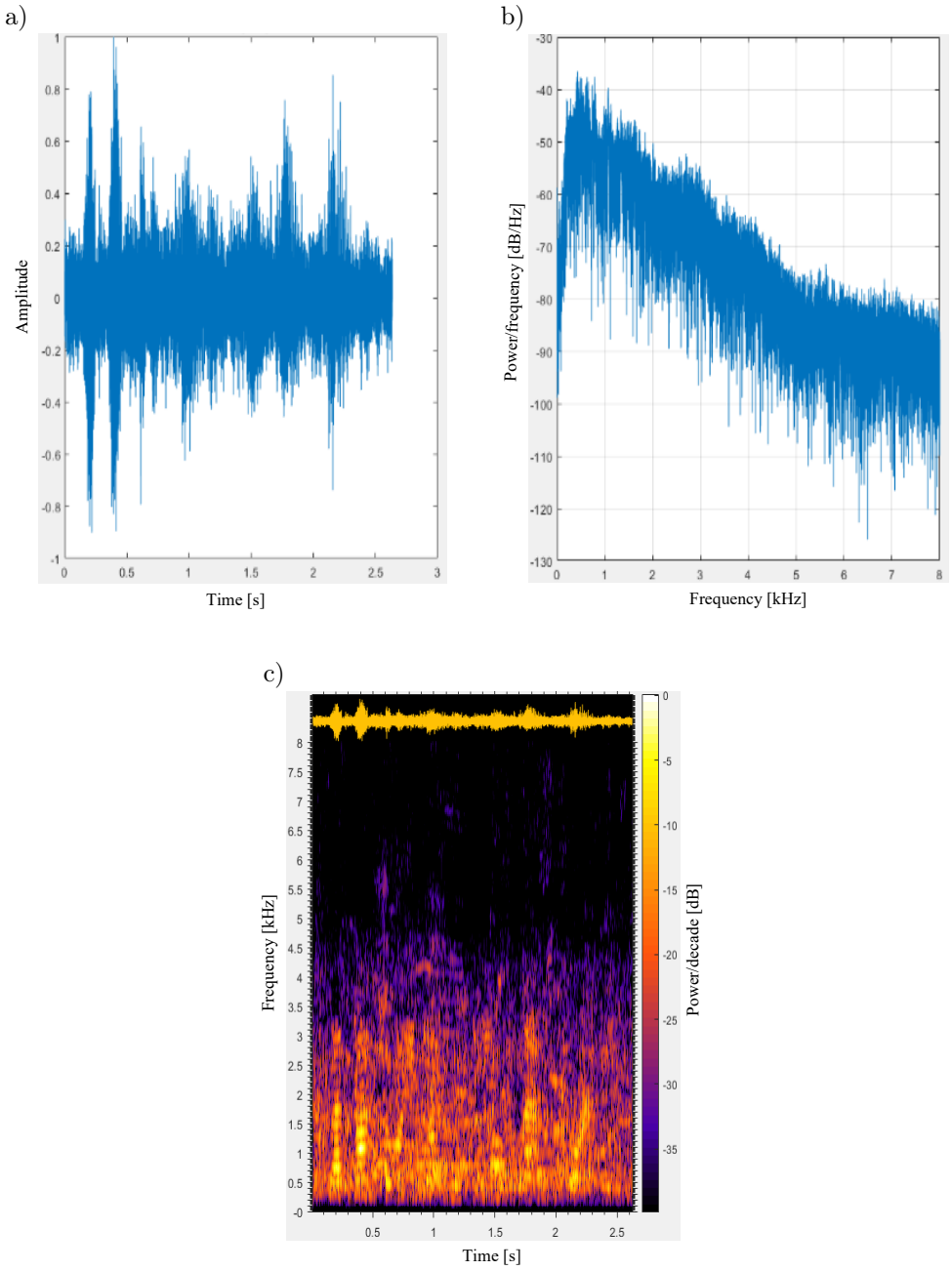


FIG. 3. Noisy speech: a) time domain, b) periodogram, and c) spectrogram.

After applying the DNN speech enhancement algorithm, the noise is removed, which can be observed in Fig. 4 that shows the denoised speech signal with

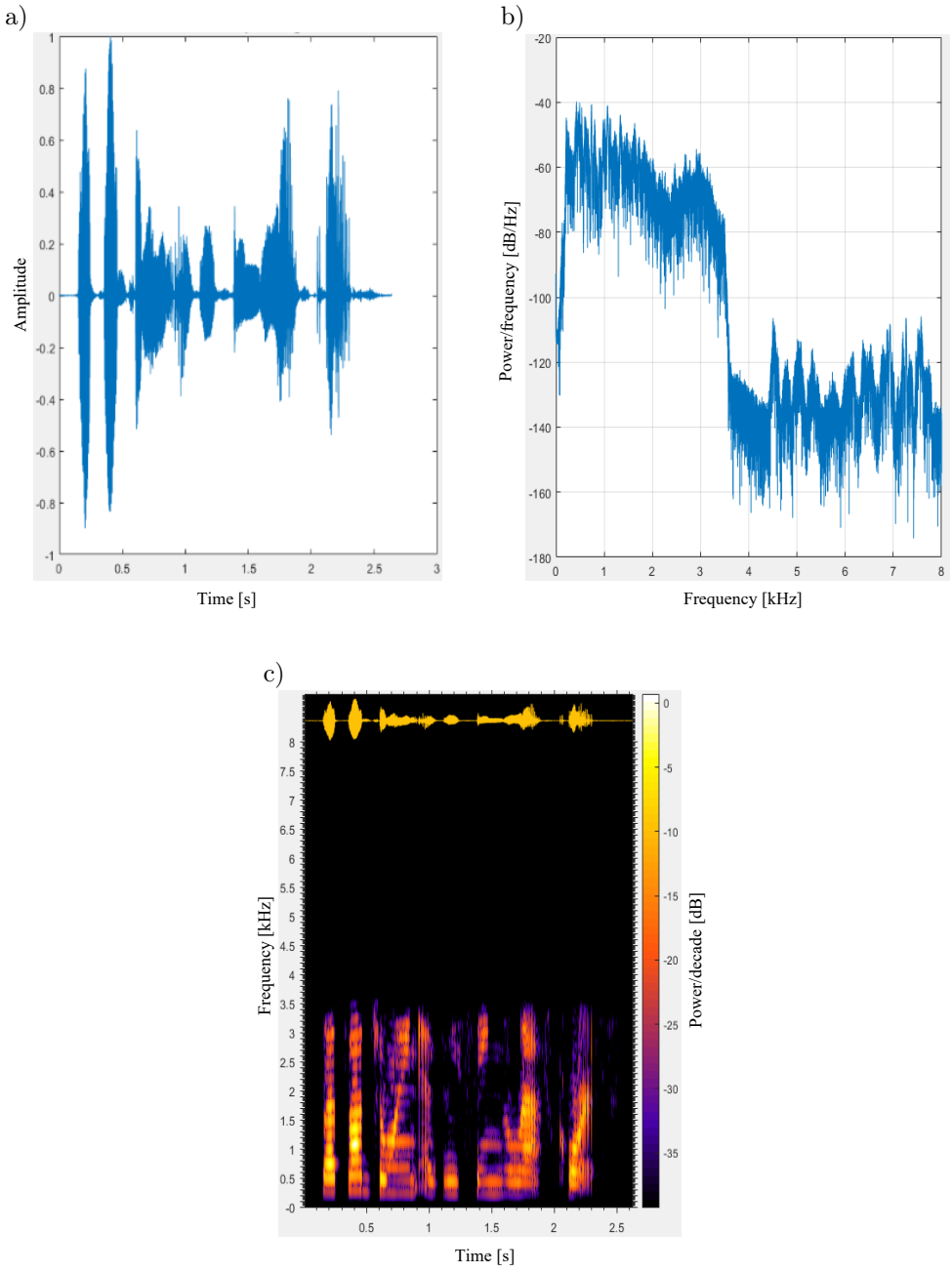


FIG. 4. Denoised speech: a) time domain, b) periodogram, and c) spectrogram.

its periodogram and spectrogram. Normalization of data helps to estimate the values between the minimum and maximum values so that it can be accessed

on a common scale. For the DNN training, the normalization of the features of the input speech signal is adjusted to zero mean and unit variance. All speech sentences are trained with the rectified linear unit. The denoised speech signal is tested for its performance based on the metrics such as the SNR and short-time objective intelligibility (STOI). STOI represents the similarity between reference and processed signal temporal envelopes for a short interval of time. STOI values are between 0 and 1, the higher values in this range indicate better intelligibility, as shown in Table 1.

Noise is removed from the noisy speech signal, and improved SNR and STOI values are shown in Table 1. Compared to the other methods adopted [48] for denoising the babble noise, the SNR is improved with this MLP DNN denoising model. The performance metrics are improved compared to the similar works adopted in speech enhancement. As the denoising system performs well for the babble noise, which is a non-stationary noise, the same methodology can be adapted for speech signals subjected to other noises for speech enhancement.

Table 1. SNR and STOI values of test sentences before and after denoising.

Input data	SNR before denoising [dB]	SNR after denoising [dB]	STOI before denoising	STOI after denoising
Test sentence 1	22.8878	27.0385	0.4915	0.6450
Test sentence 2	21.9277	27.0235	0.4014	0.5429
Test sentence 3	21.2534	26.9404	0.5544	0.5961
Test sentence 4	21.2279	27.0470	0.2677	0.4882
Test sentence 5	22.0107	27.0312	0.3192	0.5652
Test sentence 6	22.1754	26.9361	0.4983	0.6318
Test sentence 7	20.2182	26.7782	0.5444	0.6291
Test sentence 8	21.9834	26.9388	0.5175	0.6325
Test sentence 9	21.0655	27.0799	0.4337	0.6566
Test sentence 10	22.1023	27.4634	0.3488	0.5488

Figures 5 and 6 show the improvement of the denoised signal in terms of noise removal and intelligibility. The SNR of the denoised signal is more increased compared to the noisy signal, and the intelligibility of the speech signal indicates the increase in the clarity of the speech signal [48]. The denoising is very clearly observed when the denoised speech signal is listened as an audio output. The quality, as well as the intelligibility, is improved to a great extent. which shows the capability of the DNN in denoising the noisy speech and delivering the denoised signal equivalent to the clean speech signal.

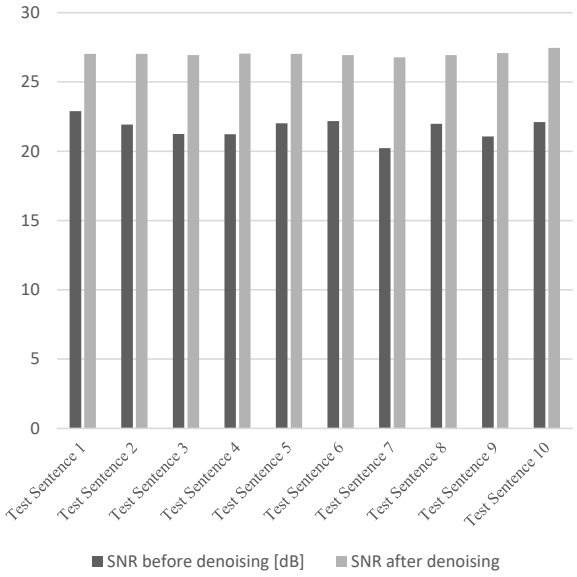


Fig. 5. The SNR of the denoised signal.

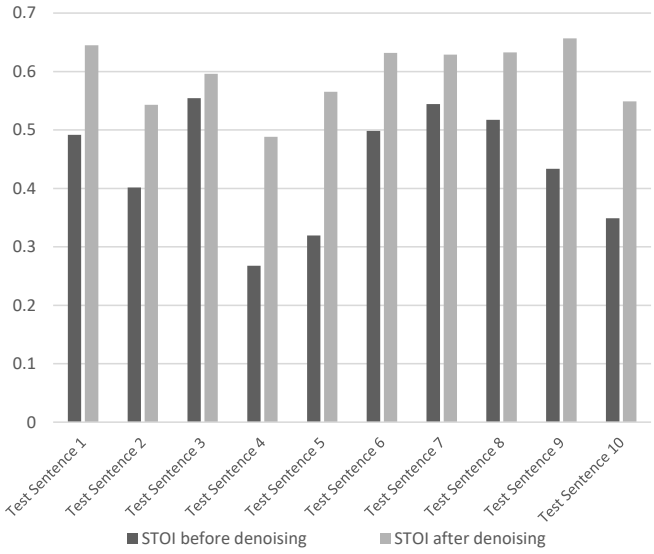


Fig. 6. STOI of noisy signal vs. denoised signal.

7. Conclusion

Background noise plays a major role in distorting the speech signal. The estimation of the ideal ratio mask yielded good results in estimating the noise and giving the denoised speech. The performance metrics such as SNR and STOI were

chosen to analyze the speech quality and intelligibility. The evaluations obtained in the performance metrics, STOI and SNR showed that the IRM-based deep learning algorithm excellently denoises the noisy speech signal and retrieves clean speech. The observations from the spectrogram also clearly indicate the removal of noise and display the denoised speech. Thus, the enhancement of speech signal was observed in the numerical values of the two-performance metrics.

References

1. J. Li, L. Deng, R. Haeb-Umbach, Y. Gong, *Robust Automatic Speech Recognition: A Bridge to Practical Applications*, 1st ed., Academic, Orlando, FL, USA, 2015.
2. B. Li, Y. Tsao, K.C. Sim, An investigation of spectral restoration algorithms for deep neural networks-based noise robust speech recognition, [in:] *Proceedings of Interspeech*, Lyon, France, pp. 3002–3006, 2013.
3. H. Levitt, Noise reduction in hearing aids: An overview, *Journal of Rehabilitation Research and Development*, **38**(1), 111–121, 2001.
4. A. Chern, Y.-H. Lai, Y.-P. Chang, Y. Tsao, R.Y. Chang, H.-W. Chang, A smartphone-based multi-functional hearing assistive system to facilitate speech recognition in the classroom, *IEEE Access*, **5**: 10339–10351, 2017, doi: 10.1109/ACCESS.2017.2711489.
5. J. Li, L. Yang, J. Zhang, Y. Yan, Comparative intelligibility investigation of single-channel noise reduction algorithms for Chinese, Japanese and English, *Journal of the Acoustical Society of America*, **129**(5): 3291–3301, 2011, doi: 10.1121/1.3571422.
6. J. Li, S. Sakamoto, S. Hongo, M. Akagi, Y. Suzuki, Two-stage binaural speech enhancement with Wiener filter for high-quality speech communication, *Speech Communication*, **53**(5): 677–689, 2011, doi: 10.1016/j.specom.2010.04.009.
7. Y. Ephraim, D. Malah, Speech enhancement using a minimum mean-square error log-spectral amplitude estimator, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **33**(2): 443–445, 1985, doi: 10.1109/TASSP.1985.1164550.
8. S. Boll, Suppression of acoustic noise in speech using spectral subtraction, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **27**(2): 113–120, Apr. 1979, doi: 10.1109/TASSP.1979.1163209.
9. Hepsiba D., J. Justin, Role of deep neural network in speech enhancement: A review, [in:] J. Hemanth, T. Silva, A. Karunananda [Eds.], *Artificial Intelligence*, SLAAI-ICAI 2018. Communications in Computer and Information Science, Vol. 890, Springer, Singapore, 2019, doi: 10.1007/978-981-13-9129-3_8.
10. P. Scalart, J.V. Filho, speech enhancement based on a priori signal to noise estimation, [in:] *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. 2, pp. 629–633, 1996, doi: 10.1109/ICASSP.1996.543199.
11. W. Xue, A.H. Moore, M. Brookes, P.A. Naylor, Modulation-domain multichannel Kalman filtering for speech enhancement, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **26**(10): 1833–1847, 2018, doi: 10.1109/TASLP.2018.2845665.
12. J. Du, Q. Huo, A speech enhancement approach using piecewise linear approximation of an explicit model of environmental distortions, [in:] *Proceedings of Interspeech*, pp. 569–572, Brisbane, Australia, 2008.

13. B. Kollmeier, R. Koch, Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction, *The Journal of the Acoustical Society of America*, **95**(3): 1593–1602, 1994, doi: 10.1121/1.408546.

14. H. Hermansky, Perceptual linear predictive (PLP) analysis of speech, *The Journal of the Acoustical Society of America*, **87**(4): 1738–1752, 1990, doi: 10.1121/1.399423.

15. H. Hermansky, N. Morgan, RASTA processing of speech, *IEEE Transactions on Speech and Audio Processing*, **2**(4): 578–589, 1994, doi: 10.1109/89.326616.

16. T. Dau, D. Püschel, A quantitative model of the "effective" signal processing in the auditory system, *The Journal of the Acoustical Society of America*, **99**(6): 3615–3622, 1996, doi: 10.1121/1.414959.

17. K. Han, Y. Wang, D.L. Wang, W.S. Woods, I. Merks, T. Zhang, Learning spectral mapping for speech dereverberation and denoising, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **23**(6): 982–992, 2015, doi: 10.1109/TASLP.2015.2416653.

18. S. Davis, P. Mermelstein, Comparison of parametric representations of monosyllabic word recognition in continuously spoken sentences, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **28**(4): 357–366, 1980, doi: 10.1109/TASSP.1980.1163420.

19. Y. Zhao, Z.-Q. Wang, D.L. Wang, Two-stage deep learning for noisy-reverberant speech enhancement, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **27**(1): 53–62, 2019, doi: 10.1109/TASLP.2018.2870725.

20. Y. Wang, A. Narayanan, D.L. Wang, On training targets for supervised speech separation, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **22**(12): 1849–1858, 2014, doi: 10.1109/TASLP.2014.2352935.

21. J. Benesty, S. Makino, J.D. Chen, *Speech Enhancement*, Springer, New York, NY, USA, 2005.

22. P.C. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, Boca Raton, FL, USA, 2013, doi: 10.1201/9781420015836.

23. H.-Y. Lee, J.-W. Cho, M. Kim, H.-M. Park, DNN-based feature enhancement using DOA-constrained ICA for robust speech recognition, *IEEE Signal Processing Letters*, **23**(8): 1091–1095, August 2016, doi: 10.1109/LSP.2016.2583658.

24. Y. Shao, S. Srinivasan, D.L. Wang, Incorporating auditory feature uncertainties in robust speaker identification, [in:] *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2007*, pp. 277–280, 2007.

25. Y. Xu, J. Du, L.-R. Dai, C.-H. Lee, A regression approach to speech enhancement based on deep neural networks, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **23**(1): 7–19, 2015, doi: 10.1109/TASLP.2014.2364452.

26. IEEE, IEEE recommended practice for speech quality measurements, *IEEE Transactions on Audio and Electroacoustics*, **17**: 225–246, 1969.

27. Y. Hu, P. Loizou, Subjective evaluation and comparison of speech enhancement algorithms, *Speech Communication*, 2007, **49**: 588–601, https://ecs.utdallas.edu/loizou/speech/noizeus/.

28. K. Tan, D. Wang, Towards model compression for deep learning based speech enhancement, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **29**: 1785–1794, 2021, doi: 10.1109/TASLP.2021.3082282.

29. F. Bao, W. Abdulla, A new ratio mask representation for CASA-based speech enhancement, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **27**(1): 7–19, 2018, doi: 10.1109/TASLP.2018.2868407.

30. Y. Liu, H. Zhang, X. Zhang, L. Yang, Supervised speech enhancement with real spectrum approximation, [in:] Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5746–5750, 2019, doi: 10.1109/ICASSP.2019.8683691.

31. C. Valentini-Botinhao, J. Yamagishi, Speech enhancement of noisy and reverberant speech for text-to-speech, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **26**(8): 1420–1433, 2018, doi: 10.1109/TASLP.2018.2828980.

32. J.-C. Hou, S.-S. Wang, Y.-H. Lai, Y. Tsao, H.-W. Chang, H.-M. Wang, Audio-visual speech enhancement using multimodal deep convolutional neural networks, *IEEE Transactions on Emerging Topics in Computational Intelligence*, **2**(20): 117–128, 2018, doi: 10.1109/TETCI.2017.2784878.

33. P. Pujol, S. Pol, C. Nadeu, A. Hagen, H. Bourlard, Comparison and combination of features in a hybrid HMM/MLP and a HMM/GMM speech recognition system, *IEEE Transactions on Speech and Audio Processing*, **13**(1): 14–22, 2005, doi: 10.1109/TSA.2004.834466.

34. Y. Xu, J. Du, L.-R. Dai, C.-H. Lee, Cross-language transfer learning for deep neural network-based speech enhancement, [in:] *Proceedings of the 9th International Symposium on Chinese Spoken Language Processing*, pp. 336–340, 2014, doi: 10.1109/ISCSLP.2014.6936608.

35. Z.-Q. Wang, D.L. Wang, Robust speech recognition from ratio masks, [in:] *Proceedings of 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5720–5724, 2016, doi: 10.1109/ICASSP.2016.7472773.

36. W. Yuan, A time–frequency smoothing neural network for speech enhancement, *Speech Communications*, **124**: 75–84, 2020, doi: 10.1016/j.specom.2020.09.002.

37. T. Lavanya, T. Nagarajan, P. Vijayalakshmi, Multi-level single channel speech enhancement using a unified framework for estimating magnitude and phase spectra, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **28**: 1315–1327, 2020, doi: 10.1109/TASLP.2020.2986877.

38. K. Sekiguchi, Y. Bando, A.A. Nugraha, K. Yoshii, T. Kawahara, Semi-supervised multichannel speech enhancement with a deep speech prior, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **27**(12): 2197–2212, 2019, doi: 10.1109/TASLP.2019.2944348.

39. F.B. Gelderblom, T.V. Tronstad, E.M. Viggen, Subjective evaluation of a noise-reduced training target for deep neural network-based speech enhancement, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **27**(3): 583–594, 2020, doi: 10.1109/TASLP.2018.2882738.

40. T. Kawase, M. Okamoto, T. Fukutomi, Y. Takahashi, Speech enhancement parameter adjustment to maximize accuracy of automatic speech recognition, *IEEE Transactions on Consumer Electronics*, **66**(2): 125–133, 2020, doi: 10.1109/TCE.2020.2986003.

41. D. Baby, T. Viratanen, J.F. Gemmeke, H. van Hamme, Coupled dictionaries for exemplar-based speech enhancement and automatic speech recognition, *IEEE/ACM Transactions*

on Audio, Speech, and Language Processing, **23**(11): 1788–1799, 2015, doi: 10.1109/TASLP.2015.2450491.

42. P.-S. Huang, M. Kim, M. Hasegawa-Johnson, P. Smaragdis, Joint optimization of masks and deep recurrent neural networks for monaural source separation, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **23**(12): 2136–2147, 2015, doi: 10.1109/TASLP.2015.2468583.

43. L. Sun, J. Du, L.-R. Dai, C.-H. Lee, Multiple-target deep learning for LSTM-RNN based speech enhancement, [in:] *2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)*, pp. 136–140, 2017, doi: 10.1109/HSCMA.2017.7895577.

44. W.-C. Huang, H.-T. Hwang, Y.-H. Peng, Y. Tsao, H.-M. Wang, Voice conversion based on cross-domain features using variational auto encoders, [in:] *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 51–55, 2018, doi: 10.1109/ISCSLP.2018.8706604.

45. W. Han, C. Wu, X. Zhang, Q. Zhang, S. Bai, Joint optimization of modified ideal ratio mask and deep neural networks for monaural speech enhancement, [in:] *Proceedings of 2017 9th International Conference on Communication Software and Networks (ICCSN)*, pp. 1070–1074, 2017, doi: 10.1109/ICCSN.2017.8230275.

46. D.S. Williamson, Y. Wang, D.L. Wang, Complex ratio masking for monaural speech separation,*IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **24**(3): 483–492, 2016, doi: 10.1109/TASLP.2015.2512042.

47. J. Ming, D. Crookes, Speech enhancement based on full-sentence correlation and clean speech recognition, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **25**(3): 531–543, 2017, doi: 10.1109/TASLP.2017.2651406.

48. R. Jaiswal, D. Romero, Implicit Wiener filtering for speech enhancement in non-stationary noise, [in:] *2021 11th International Conference on Information Science and Technology (ICIST)*, pp. 39–47, 2021, doi: 10.1109/ICIST52614.2021.9440639.

Received September 29, 2021; revised version December 15, 2021; accepted December 27, 2021.

Role of Deep Neural Network in Speech Enhancement: A Review

Chapter · July 2019

DOI: 10.1007/978-981-13-9129-3_8

2 authors, including:

Judith Justin
Avinashilingam University
21 PUBLICATIONS **60** CITATIONS

SEE PROFILE

Role of Deep Neural Network in Speech Enhancement: A Review

D. Hepsiba[1(✉)] and Judith Justin[2]

[1] Department of Instrumentation Engineering,
Karunya Institute of Technology and Sciences, Coimbatore, Tamil Nadu, India
hepsiba@karunya.edu
[2] Department of Biomedical Instrumentation Engineering,
Avinashilingam Institute for Home Science and Higher Education for Women,
Coimbatore, Tamil Nadu, India
hodbmieaul@gmail.com

Abstract. This paper presents a review on different methodologies adopted in speech enhancement and the role of Deep Neural Networks (DNN) in enhancement of speech. Mostly, a speech signal is distorted by background noise, environmental noise and reverberations. To enhance speech, certain processing techniques like Short-Time Fourier Transform, Short-time Auto-correlation and Short-time energy can be adopted. Features such as Logarithmic Power Spectrum (LPS), Mel-Frequency Cepstral Coefficients (MFCC) and Gammatone Frequency Cepstral Coefficient (GFCC) can be extracted and given to DNN for noise classification, so that the noise in the speech can be eliminated. DNN plays a major role in speech enhancement by creating a model with a large amount of training data and the performance of the enhanced speech is evaluated using certain performance metrics.

Keywords: Speech enhancement · Deep Neural Network · Feature extraction · Background noise · Speech signal

1 Introduction

Speech enhancement plays an important role in processing of any speech signal because it tends to be easily affected by different problems such as interference due to environmental noise, background noise and reverberations. Speech enhancement techniques are implemented to eliminate the environmental noise that disturbs the target speech signal and to retrieve the clean speech for applications such as Automatic Speech Recognition (ASR) [21, 22], mobile speech communication, speaker recognition, hearing aids [25, 26] and speech coding [23, 24]. Speech enhancement [1, 2] helps in improving the intelligibility and perceptual quality and also helps in the reduction of noise distortion of a speech signal degraded by adverse conditions. The different types of speech enhancement techniques developed in the past years are spectral subtraction [4], iterative wiener filtering [5], minimum mean square error [6], Kalman Filtering [15] and optimally modified log spectral amplitude [19, 20]. The presence of musical noise in the enhanced speech is the major drawback of these traditional techniques.

J. Hemanth et al. (Eds.): SLAAI-ICAI 2018, CCIS 890, pp. 103–112, 2019.
https://doi.org/10.1007/978-981-13-9129-3_8

Compared to the other traditional techniques Minimum Mean Square Error (MMSE) gives better quality of enhanced speech with lower musical noise [38, 39].

Non-linear DNN based regression models [3] are developed with training data, depending on different conditions and considering factors such as types of noise, noisy speech, noise from speakers and Signal-to-Noise Ratios (SNRs). The performance of the DNN is limited in adverse conditions and in real time noisy situations. To overcome this limitation, and to improve the generalization capability for detecting varying inputs, the training set is formed with hundreds of different noise types. This attempt proved to be efficient in managing the non-stationary behavior of noise and the different categories of unseen noise. This is done by equalizing the global variance of enhanced speech features [6] and the reference clean speech features for reducing the over-smoothing problem. The drop out training [7] is applied on the datasets of neural network when overfitting problem arises. Noise Aware Training (NAT) is done [8] by adding noise information in the DNN inputs to improve the noise robustness and performance in DNN-based speech enhancement systems.

When the need is to separate the noise from the speech signal or to separate a target source from a mixture data, the Non-negative Matrix Factorization (NMF) plays a major role. NMF has a wide scope in acoustic signal detection, speech enhancement, speech recognition in adverse environment, acoustic source separation and many more [9–12]. To increase the performance of NMF target data extraction algorithm with source subspace overlap, the estimation of encoding vectors is done by DNN to reconstruct the desired source data vectors [13]. The mixture data given to the DNN for training includes the clean speech and the noise generated from the interfering sources. DNN modeling is done by mapping the data vectors to its corresponding encoding vectors. Instead of using NMF for separation of clean speech from the mixture data, DNN can be used in two stages: first for separation of clean speech from noisy speech and second for enhancing the clean speech [17]. Another approach for enhancing speech using NMF is the exemplar-based speech enhancement technique [14], where the training clean speech and noisy data are taken in time-frequency representations. Speech and noise have varied modulation frequency content, hence, the Modulation Spectrogram feature holds good in separating the speech and noise in an efficient manner.

Time-Frequency masking is another methodology implemented when background noise causes the major problem [27]. This method improves the magnitude and phase response of the noisy speech through estimating the complex ideal ratio mask in real and imaginary domains. Here the DNN is made to learn the mapping between the reverberant speech and the complex ideal ratio mask [28].

Improved Least Mean Square Adaptive Filtering (ILMSAF) [16] helps in overcoming the drawbacks such as reduced performance in low SNR environments and poor adaptability in different noisy environments. Adaptive filter coefficients estimated by Deep Belief Network (DBN) helps in efficient noise removal. DNN acts as a noise classifier and based on the noise classification the filter parameters are chosen for removing noise.

The most commonly occurring problem in the DNN based algorithm is the reduced performance in mismatched noise condition [3, 6]. To get rid of this problem it is mandatory to have more noise types in the training set. DNN based feature extraction

can also be done to achieve speech enhancement, by learning the mapping in linear-frequency spectral domain [18]. Applying pre-enhancement in the spectral features of the DNN input could help in recovering clean speech features.

The rest of the paper is organized as follows. Section 2 discusses on the different types of databases of clean speech and noise signals. Section 3 elaborates on the processing methodologies adopted for the speech signal. Role of DNN in speech enhancement is explained in Sect. 4.

2 Databases

Database refers to both the clean speech data and noisy speech data that can be utilized for the research findings. The clean speech data is taken from the TIMIT corpus [31]. NTT database [34] has clean speech utterances in eight different languages (English, American English, Japanese, German, Chinese, Spanish, French and Italian). The DARPA-RM [29] database is suitable for training the supervised learning system. Noisy data is taken from Noisex-92 [32], Aurora-2 [30] and Speechdat-Car US (SDC) database [33]. Common noise types taken for training and testing the DNN are Babble, Restaurant, Street, Cafeteria, Machine gun, White, Volvo, Factory1, Buccaneer, etc.

3 Processing

The properties of the speech signal vary with time, and hence, the short-time processing methods that periodically repeat for the waveform duration are utilized. The following are the different processing techniques adopted in the processing of speech signal.

3.1 Short-Time Energy

Short-time energy helps in differentiating the voiced and unvoiced sounds in a speech signal. Thus, the speech and the background noise can be easily detected. The variation in short-time energy [37] determines the difference between the voiced and unvoiced speech segments. The short-time energy is high for voiced segments and low for unvoiced segments and very low for silent speech.

The short-time energy is represented as given in Eq. (1)

$$E_{\hat{n}} = \sum_{m=-\infty}^{\infty} (x[m]w[\hat{n} - m])^2 \tag{1}$$

where

$E_{\hat{n}}$-Energy of the sample n in the signal x
w-Window
m-Number of frames in the signal

3.2 Short-Time Fourier Transform

The Short-Time Fourier Transform (STFT) is the most powerful tool in any audio signal processing, especially in speech signal processing [35]. When a signal with changing frequency such as music, audio signal and speech signal is taken for noise removal, instead of analyzing the whole signal, STFT helps in analyzing the smaller divisions of the signal. The STFT is a function of both time and frequency, therefore, it is represented as time-frequency distribution [36].

The Short-Time Fourier Transform is computed using Eq. (2)

$$X[n,\lambda] = \sum_{m=-\infty}^{\infty} x[n+m]w[m]e^{-j\lambda m} \tag{2}$$

where

$n \in Z$ is a time index and $\lambda \in R$ is a normalized frequency index

3.3 Short-Time Autocorrelation

Autocorrelation is a technique that compares the original signal with the time-delayed version of itself. The Short-Time Autocorrelation is the autocorrelation function of the windowed segment of the speech signal. The voiced and unvoiced speech can be decided based on the peaks of the autocorrelation function [16].

The Short-Time Autocorrelation is denoted as given in Eq. (3)

$$R_{\hat{n}}[k] = \sum_{m=-\infty}^{\infty} (x[m]w[\hat{n}-m])(x[m+k]w[\hat{n}-k-m]) \tag{3}$$

where

$R_{\hat{n}}$ – Short-time autocorrelation at sample n in the signal x
w – Window

4 Deep Neural Networks for Speech Enhancement

DNN has wide scope in audio recognition, speech recognition, speech enhancement and other domains. DNN is a feedforward network and has the capability to model non-linear relationships. The DNN is trained with a collection of data comprising the clean and noisy speech. Different features such as Log-Power Spectra (LPS), Mel Cepstral Coefficients (MFCC) and Gammatone Frequency Cepstral Coefficients (GFCC) are extracted from the speech signal to model the DNN. During training stage, the DNN is made to learn the mapping function and the relationship between the noisy and clean speech, where the noisy data with different levels of Signal-to-Noise Ratio (SNR) are considered. In some cases, DNN is used for noise classification, where adaptive filter coefficients are selected according to the determination of noise. DNN plays a major

role in separation of the source signal from the mixed signal by decreasing the interference and distortion.

4.1 Pre-training DNN with Noisy Data

A collection of clean speech and noisy speech data represented by the log spectra features are given to the regression based DNN model in the training phase. After training, the enhanced log power spectra features are given as input to the DNN model. The DNN concatenates the time axis information in the form of multiple frames and frequency axis information in the form of log spectral features as the input feature vector for DNN learning [3].

It is observed that the performance of DNN based method gives better results compared to the logarithmic minimum mean square error (L-MMSE) method [19, 20, 40] for estimating the noise corrupted target speech. The DNN enhanced spectrogram shows no musical noise and lies closer to the original clean speech spectrogram than the L-MMSE enhanced speech. From the study made on the subjective preference evaluation, it is observed that, on an average, 76.35% of subjects have preferred DNN-based enhanced speech instead of L-MMSE enhanced speech under one or two mismatched noisy environments [3].

4.2 Drop Out Training and Noise Aware Training in DNN

The main drawback in the estimated clean speech is over-smoothing. Equalizing the global variance of estimated clean speech and reference clean speech reduces this problem to an extent. In order to remove the mismatch between the training and testing conditions caused by the different types of noise and various SNR conditions, the drop out training methodology could be adopted. Drop out Training [6] is implemented in DNN by randomly removing certain percentage of neurons from the input, intermediate or hidden layer and treated as a model. Sometimes drop out training causes decrease in performance for matched conditions but gives robustness for mismatched conditions. To give a clean picture on the noise information, Noise Aware Training [6] is done by feeding the DNN with noisy speech samples and subsequent estimation of noise. Thus, the DNN gets trained to determine the clean speech signal.

The DNN enhanced speech suppresses the non-stationary noise and results in less residual noise compared to L-MMSE enhanced speech [41]. From the study made from the subjective preference evaluation, it is observed that, on an average, 78% of the subjects have preferred DNN enhanced speech over the L-MMSE enhanced speech. It is inferred that, the DNN-based speech enhancement system is more efficient in dealing with real world noisy speech in different languages and various recording conditions that is not included in the training [6].

4.3 DNN Based Encoding Vector Estimation

Non-negative Matrix Factorization (NMF) technique is a conventional method which is used to extract encoding data vectors [9, 12]. The performance of conventional NMF based method degrades as the strength of the noise sources increase. The concept of

regression is used for estimating the encoding vectors from a mixture of data given. The mixture data and encoding vectors are mapped and learned by DNN [13].

From performance metrics such as Signal to Distortion Ratio (SDR), Signal to Interference Ratio (SIR), Signal to Artifacts Ratio (SAR) [42] and Perceptual Evaluation of Speech Quality (PESQ) [43], it is observed that the performance of DNN based NMF is good compared to the conventional NMF based techniques and DNN based separation in both matched and mismatched conditions [13].

4.4 DNN Based Noise Classification

Filter parameters play a major role in removal of noise. The filter parameters vary depending on the type of noise. DNN helps in classification of noise and selection of the filter coefficients according to the noise type. In the training phase, the Improved Least Mean Square Adaptive Filtering (ILMSAF) model is trained for different noise types. The enhancement of speech is done by selecting the ILMSAF model according to noise type. The adaptive filter coefficients play a major role in improving the perceptual quality of enhanced speech [16].

The ILMSAF based speech enhancement algorithm with DNN gives better results in terms of speech objective quality measures than the Wiener filtering method used for speech enhancement [44]. The ILMSAF based speech enhancement algorithm with DNN gives a good response in high SNR conditions and extraordinary response in low SNR conditions [16].

4.5 Source Separation and Enhancement Using DNN

The Single Channel Source Separation (SCASS) helps to separate audio source from the mixed signal [45, 47]. The most popular method is the Non-negative Matrix Factorization (NMF) and nowadays DNN is implemented for source separation also [48, 50]. Source separation is adopted by two methods using DNN. The first method maps the features of the mixed signal onto features of the source signal [49, 50]. In the second method, the spectral mask of the mixed signal is mapped, and therefore it contributes to each source in the mixed signal [51]. These methods are used for separating the sources that is distorted due to interference by other sources and distortions. Distortion is eliminated in two stages: In the first stage, the signals are denoised from the background noise, and is termed as the separation stage. Quality of the signal is enhanced in the second stage, which is the enhancement stage [17].

The separation is either done by NMF or DNN and the enhancement is done by DNN using two methods. In the first method, the separated signal is enhanced individually for each source using its own trained DNN. In the second method, a single DNN is used to enhance all the separated sources together. In both the methods, discriminative training is adopted to train the DNN in the enhancement stage. The observations made from the SIR and SDR values show that the quality of the separated sources is improved by decreasing the interference and distortions [17].

4.6 DNN Based Speech Enhancement Systems

Generally, DNN is trained in different conditions such as noise type, gender of the speaker and Signal to Noise Ratio (SNR), to ensure the generalizing capability of the DNN based speech enhancement system [53, 56] in terms of Speech Quality (SQ) and Speech Intelligibility (SI) [2]. A comparison is made in terms of noise specific, speaker specific, and signal-to-noise (SNR) specific system performance with respect to noise general, speaker general and SNR general systems. A single DNN based Speech Enhancement (SE) system has been designed for a specific noise type, speaker & SNR, is compared with the general DNN based SE system designed for various noise types, speakers & SNR and the short-time spectral amplitude minimum mean square error (STSA-MMSE) based Speech Enhancement algorithm [58].

From the performance metrics speech quality and speech intelligibility, it is observed that the DNN based SE system has good generalizing capability when exposed to unseen noise types and speakers. The DNN trained with only one type of noise, one type of speaker and one type of SNR performs excellent when compared with the general DNN based SE system trained with a variety of noise types, speakers and SNR [52].

5 Conclusion

Deep Neural Network is an emerging technique in speech enhancement and has a wide scope for research. Various processing techniques applicable for the enhancement of speech are discussed. The DNN in speech enhancement can be trained in multiple conditions and tested in mismatched conditions to test the efficiency of the network. The performance of the enhanced speech signal is evaluated with different performance metrics such as Short-Time Objective Intelligibility (STOI) score, Perceptual Evaluation of Speech Quality (PESQ), Signal to Distortion Ratio (SDR), Signal to Interference Ratio (SIR) and Signal to Artifact Ratio (SAR).

Thus, it can be concluded that the Deep Neural Network plays a major role in speech recognition, speech enhancement, audio separation and noise classification.

References

1. Benesty, J., Makino, S., Chen, J.D.: Speech Enhancement. Springer, New York, NY (2005)
2. Loizou, P.C.: Speech Enhancement: Theory and Practice. CRC Press, Boca Raton, FL (2013)
3. Xu, Y., Du, J., Dai, L.-R., Lee, C.-H.: An experimental study on speech enhancement based on deep neural networks. IEEE Signal Process. Lett. **21**(1), 65–68 (2014)
4. Boll, S.: Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. Speech Signal Process. **ASSP-27**(2), 113–120 (1979)
5. Lim, J.S., Oppenheim, A.V.: Enhancement and bandwidth compression of noisy speech. Proc. IEEE **67**(12), 1586–1604 (1979)

6. Xu, Y., Du, J., Dai, L.-R., Lee, C.-H.: A regression approach to speech enhancement based on deep neural networks. IEEE/ACM Trans. Audio Speech Lang. Process. **23**(1), 7–19 (2015)
7. Hinton, G.E., Srivastava, N., Krizheysky, A., Sutskever, I., Salakhutdinoy, R.R.: Improving neural networks by preventing co-adaptation of feature detectors (2012)
8. Seltzer, M., Yu, D., Wang, Y.: An investigation of deep neural networks for noise robust speech recognition. In: Proceedings of ICASSP, pp. 7398–7402 (2013)
9. Jin, Y.G., Kim, N.S.: On detecting target acoustic signal based on negative matrix factorization. IEICE Trans. Inf. Syst. **E93-D**(4), 922–925 (2010)
10. Wilson, K.W., Raj, B., Smaragdis, P., Divakaran, A.: Speech denoising using nonnegative matrix factorization with priors. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 4029–4032 (2008)
11. Weninger, F., Geiger, J., Wllmer, M., Schuller, B., Rigoll, G.: The Munich 2011CHiME challenge contribution: NMF-BLSTM speech enhancement and recognition for reverberated multisource environments. In: Proceedings of 1st International Workshop on Machine Listening in Multisource Environments (CHiME), pp. 24–29 (2011)
12. Grais, E.M., Erdogan, H.: Single channel speech music separation using non-negative matrix factorization and spectral masks. In: Proceedings of International Conference on Digital Signal Process, pp. 1–6 (2011)
13. Kang, T.G., Kwon, K., Shin, J.W., Kim, N.S.: NMF-based target source separation using deep neural network. IEEE Signal Process. Lett. **22**(2), 229–233 (2015)
14. Baby, D., Virtanen, T., Gemmeke, J.F., Van Hamme, H.: Coupled dictionaries for exemplar-based speech enhancement and automatic speech recognition. IEEE/ACM Trans. Audio Speech Lang. Process. **23**(11), 1788–1799 (2015)
15. Grancharov, V., Samuelsson, J., Kleijin, B.: On causal algorithms for speech enhancement. IEEE Trans. Audio Speech Lang. Process. **14**(3), 764–773 (2006)
16. Li, R., et al.: ILMSAF based speech enhancement with DNN and noise classification. Speech Commun. **85**, 53–70 (2016)
17. Grais, E.M., Roma, G., Simpson, A.J.R., Plumbley, M.D.: Two-stage single-channel audio source separation using deep neural networks. IEEE/ACM Trans. Audio Speech Lang. Process. **25**(9), 1773–1783 (2017)
18. Lee, H.-Y., Cho, J.-W., Kim, M., Park, H.-M.: DNN-based feature enhancement using DOA-constrained ICA for robust speech recognition. IEEE Signal Process. Lett. **23**(8), 1091–1095 (2016)

Cohen, I., Berdugo, B.: Speech enhancement for non-stationary noise environments. Signal Process. **81**(11), 2403–2418 (2001)

20. Cohen, I.: Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging. IEEE Trans. Speech Audio Process. **11**(5), 466–475 (2003)
21. Li, J., Deng, L., Haeb-Umbach, R., Gong, Y.: Robust Automatic Speech Recognition: A Bridge to Practical Applications, 1st edn. Academic, Orlando (2015)
22. Li, B., Tsao, Y., Sim, K.C.: An investigation of spectral restoration algorithms for deep neural networks-based noise robust speech recognition. In: Proceedings of Interspeech, pp. 3002–3006 (2013)
23. Li, J., et al.: Comparative intelligibility investigation of single-channel noise reduction algorithms for Chinese, Japanese and English. J. Acoust. Soc. Am. **129**(5), 3291–3301 (2011)
24. Li, J., Sakamoto, S., Hongo, S., Akagi, M., Suzuki, Y.: Two-stage binaural speech enhancement with Wiener filter for high-quality speech communication. Speech Commun. **53**(5), 677–689 (2011)

25. Levitt, H.: Noise reduction in hearing aids: an overview. J. Rehabil. Res. Dev. **38**(1), 111–121 (2001)
26. Chern, A., Lai, Y.H., Chang, Y.-P., Tsao, Y., Chang, R.Y., Chang, H.-W.: A smartphone-based multi-functional hearing assistive system to facilitate speech recognition in the classroom. IEEE Access **5**, 10339–10351 (2017)
27. Williamson, D.S., Wang, Y., Wang, D.: Complex ratio masking for monaural speech separation. IEEE/ACM Trans. Audio Speech Lang. Process. **24**(3), 483–492 (2016)
28. Williamson, D.S., Wang, D.L.: Time-frequency masking in the complex domain for speech dereverberation and denoising. IEEE/ACM Trans. Audio Speech Lang. Process. **25**(7), 1492–1501 (2017)
29. Price, P., Fisher, W.M., Bernstein, J., Pallet, D.: The DARPA 1000-word resource management database for continuous speech recognition. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, New York, NY, USA, pp. 651–654 (1988)
30. Hirschand, H.G.. Pearce, D.: The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions. In: Proceedings of ISCA ITRWASR, pp. 181–188 (2000)
31. Garofolo, J.S.: Getting started with the DARPA TIMIT CD-ROM: an acoustic phonetic continuous speech database. NIST Technical Report (1988)
32. Varga, A., Steeneken, H.J.M.: Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems. Speech Commun. **12**(3), 247–251 (1993)
33. Moreno et al.: Speech dat-car: a large database for automotive environments. In: Proceedings of International Conference on Language Resources and Evaluation, Athens, Greece, pp. 1–6 (2000)
34. Multi-Lingual Speech Database for Telephonometry, NTT Advanced Technology Corporation, San Jose, CA, USA (1994)
35. Allen, J.B.: Application of the short-time Fourier transform to speech processing and spectral analysis. In: Proceedings of IEEE ICASSP-82, pp. 1012–1015 (1982)
36. Cohen, L.: Time-Frequency Analysis. Englewood Cliffs, Prentice-Hall, Upper Saddle River (1995)
37. de-la-Calle-Silos, F., Stern, R.M.: Synchrony based feature extraction for robust automatic speech recognition. IEEE Signal Process. Lett. **24**(8), 1158–1162 (2017)
38. Cappe, O.: Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. IEEE Trans. Speech Audio Process. **2**(2), 345–349 (1994)
39. Hussain, A., Chetouani, M., Squartini, S., Bastari, A., Piazza, F.: Nonlinear Speech Enhancement: An Overview. In: Stylianou, Y., Faundez-Zanuy, M., Esposito, A. (eds.) Progress in Nonlinear Speech Processing. LNCS, vol. 4391, pp. 217–248. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-71505-4_12
40. Cohen, I., Gannot, S.: Spectral Enhancement Methods. In: Benesty, J., Sondhi, M.Mohan, Huang, Y.A. (eds.) Springer Handbook of Speech Processing. SH, pp. 873–902. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-49127-9_44
41. Ephraim, Y., Malah, D.: Speech enhancement using minimum mean square log spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. **ASSP-33**(2), 443–445 (1985)
42. Vincent, E., Gribonval, R., Fevotte, C.: Performance measurement in blind audio source separation. IEEE Trans. Audio Speech Lang. Process. **14**(4), 1462–1469 (2006)
43. ITU, Perceptual Evaluation of Speech Quality (PESQ): an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs ITU-T Rec. p. 862 (2000)

44. Li, R., Bao, C., Xia, B., Jia, M.: Speech enhancement using the combination of adaptive wavelet threshold and spectral sub-traction based on wavelet packet decomposition. In: 2012 IEEE 11th International Conference on Signal Processing (ICSP), vol. 1, pp. 481–484 (2012)
45. Virtanen, T.: Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria. IEEE Trans. Audio Speech Lang. Process. **15**(3), 1066–1074 (2007)
46. Smaragdis, P.: Convolutive speech bases and their application to supervised speech separation. IEEE Trans. Audio Speech Lang. Process. **15**(1), 1–12 (2007)
47. Smaragdis, P., Shashanka, M., Raj, B.: A sparse non-parametric approach for single channel separation of known sounds. In: Neural Information Processing Systems, Vancouver, BC, Canada, Dec 2009, pp. 1705–1713
48. Grais, E.M., Roma, G., Simpson, A.J.R., Plumbley, M.D.: Single channel audio source separation using deep neural network ensembles. In: Proceedings of 140th Audio Engineering Society Convention, Paper no. 9494 (2016)
49. Huang, P.-S., Kim, M., Hasegawa-Johnson, M., Smaragdis, P.: Singing-voice separation from monaural recordings using deep recurrent neural networks. In: Proceedings of International Society for Music Information Retrieval Conference, pp. 477–482 (2014)
50. Huang, P.-S., Kim, M., Hasegawa-Johnson, M., Smaragdis, P.: Deep learning for monaural speech separation. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 1562–1566 (2014)
51. Weninger, F., Hershey, J.R., Roux, J.L., Schuller, B.: Discriminatively trained recurrent neural networks for single-channel speech separation. In: Proceedings of IEEE Global Conference on Signal and Information Processing, pp. 577–581 (2014)
52. Kolbæk, M., Tan, Z.-H.: Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems. IEEE/ACM Trans. Audio Speech Lang. Process. **25**(1), 153–167 (2017)
53. Lee, T., Theunissen, F.: A single microphone noise reduction algorithm based on the detection and reconstruction of spectro-temporal features. Proc. R. Soc. Lond. A Math. Phys. Sci. **471**, 2184 (2015)
54. Huang, P.-S., Kim, M., Hasegawa-Johnson, M., Smaragdis, P.: Joint optimization of masks and deep recurrent neural networks for monaural source separation. IEEE/ACM Trans. Audio Speech Lang. Process. **23**(12), 2136–2147 (2015)
55. Liu, D., Smaragdis, P., Kim, M.: Experiments on deep learning for speech denoising. In: Proceedings of INTERSPEECH, pp. 2685–2689 (2014)
56. Wang, Y., Chen, J., Wang, D.: Deep neural network based supervised speech segregation generalizes to novel noises through large-scale training. Dept. Comput. Sci. Eng. Ohio State Univ., Columbus, OH, USA, Technical Report OSU-CISRC-3/15-TR02 (2015)
57. Hendriks, R.C., Gerkmann, T., Jensen, J.: DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement: A Survey of the State of the Art. Synthesis Lectures on Speech and Audio Processing, vol. 9, pp. 1–80. Morgan & Claypool, SanRafael, CA (2013)
58. Erkelens, J., Hendriks, R., Heusdens, R., Jensen, J.: Minimum mean square error estimation of discrete Fourier coefficients with generalized Gamma priors. IEEE/ACM Trans. Audio Speech Lang. Process. **15**(6), 1741–1752 (2007)