CHAPTER 6
ALARYNGEAL SPEECH ENHANCEMENT

6.1 CHANGES IN VOICE CAUSED DUE TO LARYNGECTOMY

Sound is produced when air from the lungs passes through the vocal cords in the larynx, causing them to vibrate. The vibration of the vocal cords produces sound, which is then shaped into speech by the mouth, tongue, and lips. A laryngectomy is a surgical procedure that involves the removal of the larynx. This surgery is most often performed to treat advanced laryngeal cancer but may also be done due to severe injury or other medical conditions. After this procedure, a person will no longer be able to speak in the traditional way and will breathe through a permanent opening in the neck called a stoma. Patients often use alternative communication methods, such as esophageal speech, a voice prosthesis, or an electrolarynx. A small hole or puncture is made in the shared wall between the trachea and oesophagus. This puncture is visible inside the stoma and becomes a path to allow airflow into the oesophagus. A small silicone one-way valve is placed inside the puncture to keep it open. This one-way valve is called a tracheoesophageal prosthesis. The speech uttered by a person who has undergone total laryngectomy patients is called alaryngeal speech (Aguilar-Torres et al., 2006).; Doi et al., (2014). This chapter deals with a comparative evaluation of performance metrics of deep learning algorithms for enhanced alaryngeal speech.

6.2 TYPES OF SPEECH ENHANCEMENT

The reduction of different types of background noise at various decibels encompasses the types of speech enhancement, such as near-end and far-end speech enhancement, depending on the context in which the noise reduction is being applied.

When focusing on reducing background noise for the speech signal captured by a microphone close to the speaker (near end), it falls under the category of near-end speech enhancement. Near-end speech enhancement represents the addition of noise to the signal during perception, though the signal is generated through a clean environment. Example: Railway Announcement system – though generated appropriately, the listener finds it difficult to perceive due to additive noise in the perceiving end.

Far end Speech is a speech transmitted through a channel, and noise gets added to the channel, and the receiver receives only noise speech. In this case, the focus is on improving the quality of the speech signal the listener receives (far end), which may involve reducing noise that was added or increased during transmission. Far-end speech enhancement works on the signal received by the listener's end after it has traveled through a communication medium.

6.3 SPEECH ENHANCEMENT SYSTEM FOR ALARYNGEAL SPEECH

Figure 6.1 shows the speech enhancement system that helps to improve the quality of alaryngeal speech.

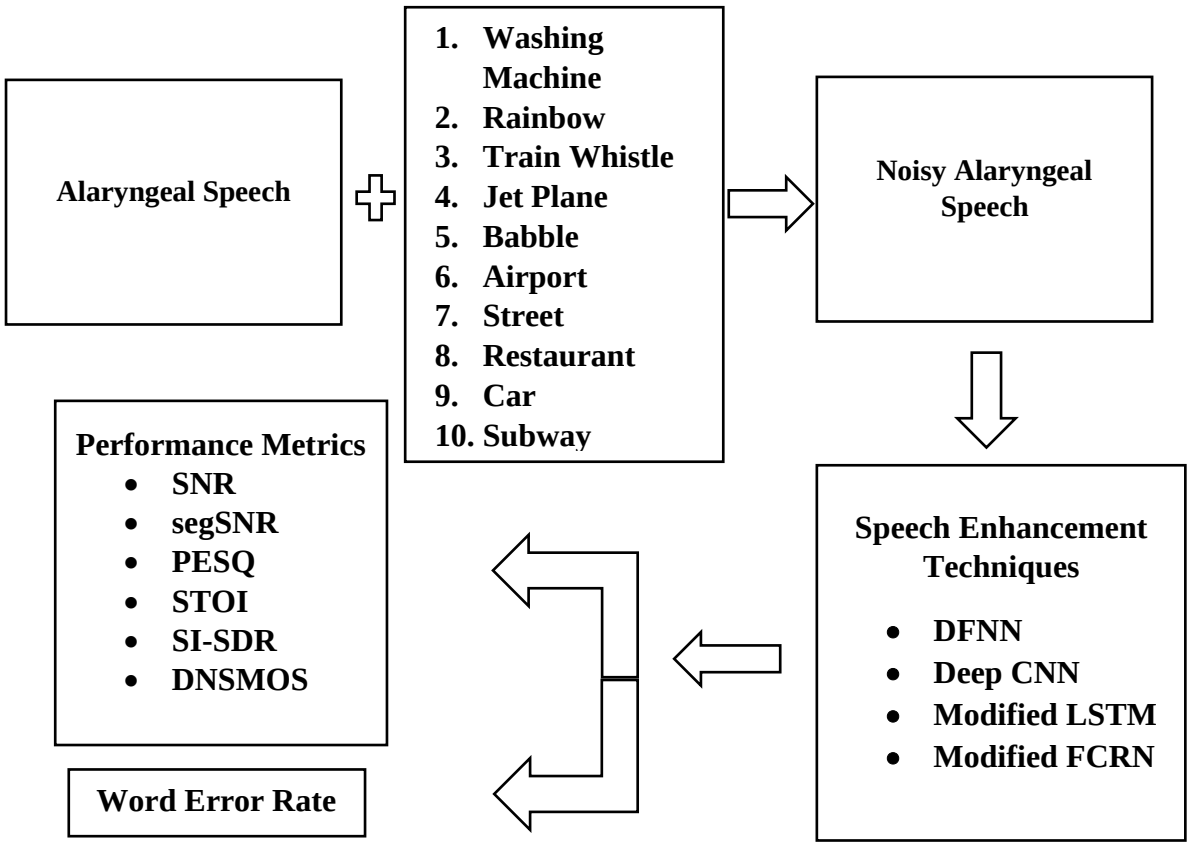


Figure 6.1 Speech Enhancement System for Alaryngeal Speech

Due to this reason, it is challenging for others to understand the speech spoken by total laryngectomy patients with Blom Singer non-indwelling voice prosthesis. Various speech enhancement methodologies are adopted to enhance the alaryngeal speech subjected to various types of noises in real time Aguilar-Torres et al., (2006).

6.4 DATASET FOR ALARYNGEAL SPEECH

The alaryngeal speech dataset is created by recording the Harvard sentences such as “Read verse out loud for pleasure” and “The stray cat gave birth to kittens” spoken by the patient who has undergone total laryngectomy. The Harvard sentences, or Harvard lines (“IEEE Recommended Practice for Speech Quality Measurements,” 1969), is a collection of 720 sample phrases divided into lists of 10 phonetically balanced sentences that use specific phonemes at the same frequency. Clean alaryngeal speech is mixed with different noise types, such as Washing Machine Noise, Rainbow Noise, Train whistle Noise, Jet Airplane Noise, Babble Noise, Street Noise, Airport Noise, Restaurant Noise, Car Noise, and Subway Noise at different noise levels, such as -10dB, -5dB, 0dB, 5dB, 10dB, and 15dB. The different types of noises are taken from MATLAB\R2020b\ toolbox\audio\samples and https://www.ee.columbia.edu/~dpwe/sounds/noise/ to simulate real world conditions. For training and testing, a variety of noise types are incorporated, including washing machine noise, rainbow noise, jet airplane noise, and train whistle noise from MATLAB\R2020b\ toolbox\audio\samples, babble noise, airport noise, street noise, and restaurant noise sourced from the Columbia database. To assess the performance of the speech enhancement system, unseen noises such as car and subway noises are taken from the Columbia database.

Alaryngeal speech recordings were obtained from 10 laryngectomees, consisting of 9 males and 1 female, aged between 35 and 60 years. These recordings were made in an anechoic chamber using a Condenser Studio XLR microphone with echo cancellation at a 16 kHz sampling rate. The data thoroughly examines alaryngeal speech characteristics, aiding in detailed analysis and modeling in speech processing research. The microphone was positioned 10 centimeters from the speakers to ensure optimal signal integrity. The recordings feature phonetically balanced sentences from the renowned Harvard Sentences corpus, selected for their ease of utterance by the patients. Sentences like “Read verse out loud for pleasure” and “The stray cat gave birth to kittens” were chosen. Participants who faced challenges in producing consistent sound sipped water to clear their throats before speaking. Each sentence was repeated two to three times for comfort, resulting in a total of 54 recordings.

6.5 ALGORITHMS TESTED FOR SPEECH ENHANCEMENT

The noisy alaryngeal speech is generated by adding Washing Machine noise, Rainbow noise, Train whistle noise, Jet Airplane noise, Babble noise, Street noise, Airport noise, Restaurant noise, Car noise, and Subway noise at different noise levels of -10dB, -5dB, 0dB, 5dB, 10dB, and 15dB. The different algorithms, such as DFNN, Deep CNN, modified LSTM, and modified FCRN, were trained and tested with alaryngeal speech for speech enhancement. The performance of the speech enhancement is analyzed based on performance metrics such as SNR, segSNR, PESQ, STOI, SI-SDR, and DNSMOS, and the results are given in tables 6.1 to 6.6. Spectrograms are shown in Figures 6.2 to 6.4. Speech Quality and Intelligibility metrics are represented in Figures 6.5 to 6.10.

6.6 RESULTS AND DISCUSSION

Based on the performance evaluation of enhanced alaryngeal speech given in Tables 6.1 to 6.6, it is evident that the modified FCRN technique has performed well in terms of the performance metrics for improving the quality and intelligibility of alaryngeal speech.

The performance of the modified FCRN technique in terms of SNR observed in Figure 6.2 shows notable improvement across various noise types, including jet plane noise, train whistle noise, washing machine noise, rainbow noise, and airport noise. The results indicate that the modified LSTM technique exhibit commendable performance, whereas Deep CNN demonstrate superior noise removal capabilities compared to DFNN. Overall, the modified FCRN technique effectively enhances SNR, with deep CNN proving to be the most effective in reducing noise among the evaluated approaches.

The segSNR values in Figure 6.3 reveal that the modified FCRN algorithm significantly improves noise reduction, particularly for noise types such as rainbow, babble, jet plane, and street. The modified FCRN technique stands out for its superior ability to enhance speech signal quality when compared across various noise types and levels. It outperforms other algorithms, including modified LSTM, deep CNN, and DFNN, demonstrating its effectiveness in improving segSNR and overall speech clarity.

PESQ values observed in Figure 6.4 indicate substantial improvements across all noise types and levels, with the modified FCRN technique achieving notably high-quality

scores. The highest PESQ ratings are observed for train whistle noise, babble noise, street noise, and rainbow noise, where the speech quality falls within the "Excellent Quality" range. For other noise types, the PESQ scores are categorized as "Good," reflecting the effective enhancement of speech signal quality by the modified FCRN technique. Overall, the modified FCRN approach performs better in improving speech quality than other algorithms, with consistently high PESQ ratings indicating significant advancements in perceptual speech quality.

Figure 6.5 demonstrates that the modified FCRN technique significantly enhances speech intelligibility across various noise types and levels. The STOI scores for the modified FCRN approach are notably higher than those of other algorithms, which exhibit lower intelligibility scores. This indicates that modified FCRN performs better in maintaining and improving speech clarity and understanding, as evidenced by its consistently higher STOI values.

The evaluation of SI-SDR in Figure 6.6 indicates that the modified FCRN technique excels in noise reduction across various noise types at different noise levels. The modified FCRN approach consistently achieves superior SI-SDR scores compared to other algorithms, such as modified LSTM, Deep CNN, and DFNN, demonstrating its effectiveness in improving signal quality and minimizing distortion. While the modified FCRN technique delivers outstanding performance, the other algorithms show comparatively lower SI-SDR values, highlighting the modified FCRN's advanced capability in enhancing overall noise reduction.

The values of DNSMOS in Figure 6.7 highlight that the modified FCRN technique significantly enhances noise suppression quality, as evidenced by the high DNSMOS scores achieved for various noise types and noise levels. Specifically, the modified FCRN approach demonstrates superior performance in noise suppression compared to other algorithms, with notably higher DNSMOS scores indicating effective subjective quality improvement. This performance underscores modified FCRN's capability to deliver exceptional noise suppression and superior quality when evaluating noise suppression algorithms.

In handling unseen noises, the DFNN is the least effective in addressing the complexities of unseen noises like car or subway sounds. However, both Deep CNN and the modified LSTM demonstrate a moderate level of competence, as indicated by the

performance metrics. Modified FCRN shows outstanding performance in terms of handling both unseen noises.

From the values of performance metrics such as SNR, segSNR, SI-SDR, PESQ, STOI, and DNSMOS, it is evident that the modified FCRN technique performs well in improving the quality and intelligibility of alaryngeal speech.

Table 6.1 Performance Evaluation of Enhanced Alaryngeal Speech at -10 dB Noise level

Noise Type	SNR				segSNR				PESQ				STOI				SI-SDR				DNSMOS			
	DFNN	Deep CNN	Modified LSTM	Modified FCRN	DFNN	Deep CNN	Modified LSTM	Modified FCRN	DFNN	Deep CNN	Modified LSTM	Modified FCRN	DFNN	Deep CNN	Modified LSTM	Modified FCRN	DFNN	Deep CNN	Modified LSTM	Modified FCRN	DFNN	Deep CNN	Modified LSTM	Modified FCRN
Washing Machine	24.86	29.31	30.54	31.75	3.95	5.02	6.11	7.55	1.62	1.78	2.18	3.14	0.32	0.36	0.43	0.73	3.56	5.04	5.64	9.11	2.32	2.57	2.82	3.18
Rainbow	25.14	28.36	29.05	31.65	4.47	5.7	6.93	7.99	1.36	1.45	1.82	3.21	0.35	0.49	0.52	0.69	3.44	4.76	5.47	9.45	2.44	2.72	2.85	3.28
Babble	28.89	33.03	33.81	34.23	3.73	4.83	6.07	8.14	1.54	1.96	2.24	3.47	0.39	0.48	0.52	0.78	3.72	4.93	5.46	8.92	2.39	2.79	2.9	3.28
Airport	29.48	32.1	36.14	37.14	4.05	5.61	6.81	8.2	1.54	1.87	2.21	3.31	0.42	0.47	0.55	0.77	3.56	4.82	5.43	9.24	2.42	2.66	2.93	3.33
Jet Plane	30.27	32.45	35.44	36.62	4.01	5.35	6.56	8.22	1.37	1.86	2.2	3.21	0.43	0.49	0.57	0.71	3.69	4.99	5.59	9.33	2.35	2.84	3.05	3.35
Street	27.22	31.19	32.47	32.74	3.15	4.59	6.03	8.33	1.46	1.69	2.09	3.11	0.47	0.49	0.53	0.75	3.53	4.81	5.42	8.11	2.44	2.81	2.94	3.3
Train Whistle	32.48	33.12	34.79	36.51	3.22	4.36	5.66	8.24	1.58	1.78	1.95	2.95	0.51	0.54	0.62	0.78	3.36	4.64	5.22	9.22	2.41	2.84	2.95	3.42
Restaurant	31.69	31.46	32.78	34.2	3.95	5.19	6.66	8.12	1.42	1.87	2.06	2.88	0.41	0.43	0.49	0.76	3.28	4.58	5.16	9.71	2.44	2.87	3.05	3.67
Car	25.64	29.53	30.24	33.63	2.55	4.02	5.4	7.28	1.29	1.73	2.05	2.92	0.31	0.44	0.51	0.73	2.89	4.42	5.03	8.56	2.18	2.69	2.87	3.38
Subway	26.47	28.74	30.84	32.88	2.82	4.39	5.71	7.86	1.36	1.77	2.08	3.04	0.33	0.46	0.54	0.76	2.98	4.56	5.18	8.96	2.23	2.72	2.93	3.56

Table 6.2 Performance Evaluation of Enhanced Alaryngeal Speech at -5 dB Noise level

Noise Type	SNR				segSNR				PESQ				STOI				SI-SDR				DNSMOS			
	DFNN	Deep CNN	Modified LSTM	Modified FCRN	DFNN	Deep CNN	Modified LSTM	Modified FCRN	DFNN	Deep CNN	Modified LSTM	Modified FCRN	DFNN	Deep CNN	Modified LSTM	Modified FCRN	DFNN	Deep CNN	Modified LSTM	Modified FCRN	DFNN	Deep CNN	Modified LSTM	Modified FCRN
Washing Machine	27.47	31.26	33.05	34.62	5.68	6.71	7.9	9.21	1.67	1.86	2.29	3.45	0.37	0.41	0.48	0.78	5.12	5.65	6.86	11.05	2.57	2.87	3.26	3.63
Rainbow	28.21	31.59	30.49	33.5	7	8.25	9.35	9.1	1.53	1.62	2.34	3.46	0.41	0.53	0.59	0.77	5.06	5.37	6.36	11.46	2.63	2.89	2.97	3.51
Babble	30.95	32.82	34.48	35.81	6.09	7.39	8.69	9.32	1.65	2.18	2.5	3.71	0.43	0.51	0.6	0.83	5.33	5.29	6.49	10.95	2.65	2.91	3.19	3.43
Airport	30.65	33.77	36.56	38.26	6.55	8.27	9.5	9.33	1.72	1.91	2.54	3.49	0.48	0.54	0.62	0.82	4.87	5.13	6.32	10.84	2.66	2.91	3.19	3.45
Jet Plane	32.57	33.29	37.13	37.97	6.74	8.15	9.41	9.76	1.54	2.08	2.32	3.46	0.48	0.57	0.62	0.79	5.01	5.3	6.12	11.38	2.52	3.03	3.38	3.41
Street	30.44	32.78	33.65	34.16	5.75	7.09	8.74	9.48	1.54	1.75	2.48	3.34	0.55	0.56	0.61	0.79	4.84	5.15	6.29	10.83	2.59	2.96	3.03	3.4
Train Whistle	33.02	33.66	35.65	37.05	5.26	6.04	7.29	9.33	1.76	1.96	2.26	3.46	0.58	0.61	0.67	0.83	4.67	4.95	6.11	11	2.56	2.98	3.01	3.56
Restaurant	33.87	32.11	34.05	35.08	5.55	6.79	8.37	9.22	1.53	1.94	2.27	3.13	0.44	0.49	0.52	0.81	4.47	4.89	6.57	11.31	2.57	3	3.21	3.85
Car	27.259	31.61	32.79	34.91	4.25	5.7	7.03	8.94	1.34	1.95	2.21	3.37	0.38	0.51	0.56	0.82	4.37	5.76	5.85	10.54	2.36	2.84	3.15	3.53
Subway	28.884	30.05	32.06	34.2	4.86	6.13	8.42	9.04	1.48	1.86	2.24	3.42	0.37	0.53	0.61	0.83	4.49	5.9	6.17	11.33	2.47	2.89	3.28	3.74

Table 6.3 Performance Evaluation of Enhanced Alaryngeal Speech at 0 dB Noise level

Noise Type	SNR				segSNR				PESQ				STOI				SI-SDR				DNSMOS			
	DFNN	Deep CNN	Modified LSTM	Modified FCRN	DFNN	Deep CNN	Modified LSTM	Modified FCRN	DFNN	Deep CNN	Modified LSTM	Modified FCRN	DFNN	Deep CNN	Modified LSTM	Modified FCRN	DFNN	Deep CNN	Modified LSTM	Modified FCRN	DFNN	Deep CNN	Modified LSTM	Modified FCRN
Washing Machine	29.59	33.37	34.84	37.46	7.58	8.54	9.98	11.01	1.81	2.11	2.48	3.77	0.45	0.45	0.58	0.83	6.69	7.85	9.23	13.27	2.71	3.14	3.48	3.84
Rainbow	29.92	33.49	34.72	35.84	8.52	9.6	10.64	10.72	1.62	1.7	2.43	3.79	0.46	0.62	0.68	0.82	6.83	7.57	8.76	13.31	2.74	3.05	3.2	3.72
Babble	31.07	34.85	35.54	36.91	7.64	8.82	10.23	10.95	1.83	2.26	2.72	4.01	0.46	0.55	0.62	0.86	7.1	7.76	8.85	12.04	2.81	3.12	3.45	3.52
Airport	31.84	34.91	38.79	40.28	8.02	9.96	11.05	10.97	1.96	2.2	2.72	3.77	0.52	0.58	0.79	0.85	6.6	7.63	8.61	13.07	2.61	3.16	3.51	3.63
Jet Plane	32.18	34.38	38.06	38.83	8.17	9.62	10.54	11.05	1.61	2.15	2.43	3.78	0.53	0.64	0.67	0.82	6.8	7.87	8.84	13.05	2.82	3.16	3.65	3.8
Street	29.19	33.04	34.83	35.87	7.47	8.99	10.12	11.16	1.64	1.91	2.64	3.78	0.58	0.64	0.66	0.83	6.62	7.7	8.65	13.07	2.72	3.22	3.18	3.63
Train Whistle	33.25	35.59	36.74	38.92	6.29	7.38	8.43	11.17	1.93	2.19	2.41	3.79	0.62	0.68	0.72	0.85	6.39	7.44	8.39	13.29	2.69	3.29	3.19	3.73
Restaurant	33.94	34.57	35.68	37.75	6.84	8.06	9.4	10.97	1.64	2.1	2.39	3.56	0.49	0.53	0.59	0.85	6.25	7.39	8.44	13.44	2.75	3.08	3.46	3.97
Car	29.7	33.16	35.21	36.63	5.63	6.83	8.17	10.74	1.52	2.03	2.42	3.58	0.43	0.59	0.64	0.84	5.36	7.31	8.14	11.33	2.59	3.12	3.39	3.84
Subway	30.58	33.41	34.53	37.08	5.85	7.59	9.8	10.67	1.64	2.06	2.49	3.81	0.41	0.6	0.69	0.86	5.21	7.49	8.47	13.18	2.62	3.19	3.49	3.91

Table 6.4 Performance Evaluation of Enhanced Alaryngeal Speech at 5 dB Noise level

Noise Type	SNR				segSNR				PESQ				STOI				SI-SDR				DNSMOS			
	DFNN	Deep CNN	Modified LSTM	Modified FCRN	DFNN	Deep CNN	Modified LSTM	Modified FCRN	DFNN	Deep CNN	Modified LSTM	Modified FCRN	DFNN	Deep CNN	Modified LSTM	Modified FCRN	DFNN	Deep CNN	Modified LSTM	Modified FCRN	DFNN	Deep CNN	Modified LSTM	Modified FCRN
Washing Machine	34.09	35.47	36.49	39.35	9.32	10.39	11.62	12.95	2.02	2.02	2.75	4.09	0.55	0.54	0.68	0.862	9.76	11.24	12.64	15.34	2.85	3.53	3.8	4.04
Rainbow	32.57	34.41	37.72	38.38	9.89	11.08	12.48	12.44	1.78	2.07	2.89	4.01	0.56	0.62	0.66	0.86	9.69	10.96	11.97	15.63	3.05	3.28	3.51	3.96
Babble	32.84	34.71	36.37	37.7	8.97	10.28	11.59	12.67	2.04	2.51	2.95	4.3	0.6	0.63	0.68	0.87	9.91	11.17	12.42	14.34	3.02	3.49	3.75	3.87
Airport	33.15	36.02	39.06	41.57	9.25	11.15	12.49	12.68	1.94	2.57	3.17	4.08	0.62	0.65	0.82	0.89	9.75	11.02	11.68	15.16	3.17	3.38	3.84	3.91
Jet Plane	34.03	36.58	38.78	39.7	9.8	11.45	12.73	12.77	1.77	2.36	2.99	4.09	0.64	0.72	0.76	0.84	9.97	11.03	11.93	15.39	2.96	3.48	3.85	4.03
Street	32.03	37.09	37.15	37.39	8.66	10.15	11.75	12.99	1.81	2.37	3.42	4.19	0.72	0.77	0.82	0.88	9.77	10.85	11.72	15.24	2.98	3.51	3.48	3.85
Train Whistle	35.86	36.61	37.48	38.04	7.9	9.39	10.18	12.89	1.91	2.49	2.87	4.14	0.71	0.75	0.82	0.87	9.54	10.83	11.46	15.49	2.95	3.52	3.49	3.85
Restaurant	35.13	35.29	36.61	38.84	8.17	9.41	11.19	12.59	1.79	2.46	2.76	3.97	0.59	0.62	0.71	0.86	9.53	10.87	11.42	15.63	3.04	3.28	3.85	4.04
Car	31.258	34.21	35.94	37.75	7.26	9.05	9.92	12.68	1.64	2.28	2.67	3.84	0.52	0.67	0.72	0.86	8.22	11.46	11.64	12.85	2.8	3.39	3.74	3.99
Subway	31.556	34.47	36.48	38.62	7.53	9.21	11.43	12.39	1.79	2.19	2.75	4.12	0.54	0.69	0.76	0.89	8.86	11.92	11.78	15.86	2.98	3.27	3.85	4.21

Table 6.5 Performance Evaluation of Enhanced Alaryngeal Speech at 10 dB Noise level

Noise Type	SNR				segSNR				PESQ				STOI				SI-SDR				DNSMOS			
	DFNN	Deep CNN	Modified LSTM	Modified FCRN	DFNN	Deep CNN	Modified LSTM	Modified FCRN	DFNN	Deep CNN	Modified LSTM	Modified FCRN	DFNN	Deep CNN	Modified LSTM	Modified FCRN	DFNN	Deep CNN	Modified LSTM	Modified FCRN	DFNN	Deep CNN	Modified LSTM	Modified FCRN
Washing Machine	35.72	37.92	38.61	41.28	10.87	11.87	13.21	14.25	2.06	2.38	2.94	4.21	0.62	0.65	0.75	0.91	11.47	13.78	15.76	16.62	3.46	3.95	4.23	4.42
Rainbow	35.59	36.86	40.15	40.71	10.87	12.28	13.84	14.16	1.98	2.44	3.24	4.21	0.65	0.69	0.74	0.93	11.4	13.5	15.9	17.83	3.24	3.54	3.84	4.11
Babble	35.29	37.26	38.49	40.39	9.6	11.05	12.5	14.47	2.19	2.88	3.55	4.41	0.71	0.74	0.75	0.92	11.68	13.77	15.97	16.62	3.31	3.79	4.29	4.32
Airport	34.8	37.39	41.52	42.83	10.16	12.14	13.55	14.49	2.54	2.77	3.8	4.11	0.67	0.72	0.85	0.93	11.73	13.61	15.86	16.98	3.29	3.52	4.19	4.36
Jet Plane	37.35	39.41	41.03	44.12	10.74	12.41	14.21	14.49	1.86	2.89	3.69	4.19	0.7	0.8	0.85	0.88	11.92	12.94	16.07	16.67	3.37	3.75	4.08	4.21
Street	34.21	36.77	38.26	39.39	10.21	11.99	13.59	14.67	2	2.64	3.85	4.37	0.77	0.8	0.85	0.92	12.7	12.75	15.87	17.38	3.35	3.58	3.82	4.13
Train Whistle	36.23	38.19	39.39	41.07	8.98	9.51	11.51	14.54	2.51	2.77	3.4	4.35	0.8	0.81	0.85	0.89	11.52	13.42	15.64	17.54	3.32	3.56	3.83	4.12
Restaurant	37.75	39.06	39.94	41.97	9.26	10.65	12.54	14.27	2.28	2.89	3.36	4.11	0.72	0.76	0.77	0.91	11.73	13.42	15.57	17.75	3.41	3.4	4.22	4.32
Car	32.841	36.82	39.58	40.34	9.43	10.17	11.36	13.98	1.83	2.65	3.14	4.21	0.61	0.73	0.78	0.89	10.29	12.36	14.8	15.24	3.21	3.27	4.08	4.13
Subway	32.657	36.83	38.84	40.98	9.58	11.26	13.27	14.19	1.95	2.73	3.29	4.34	0.65	0.75	0.81	0.9	10.94	13.54	14.91	17.25	3.1	3.41	4.17	4.37

Table 6.6 Performance Evaluation of Enhanced Alaryngeal Speech at 15 dB Noise level

Noise Type	SNR				segSNR				PESQ				STOI				SI-SDR				DNSMOS			
	DFNN	Deep CNN	Modified LSTM	Modified FCRN	DFNN	Deep CNN	Modified LSTM	Modified FCRN	DFNN	Deep CNN	Modified LSTM	Modified FCRN	DFNN	Deep CNN	Modified LSTM	Modified FCRN	DFNN	Deep CNN	Modified LSTM	Modified FCRN	DFNN	Deep CNN	Modified LSTM	Modified FCRN
Washing Machine	36.02	38.12	41.7	43.5	14.18	15.07	16.42	17.24	2.31	2.9	3.62	4.52	0.73	0.81	0.86	0.93	16.4	17.4	20.04	20.48	3.68	4.24	4.63	4.7
Rainbow	35.91	37.34	42.73	43.49	14.63	16.04	17.64	18.19	1.95	2.7	4.11	4.55	0.66	0.71	0.77	0.95	16.33	17.12	20.11	21.02	3.38	3.73	4.53	4.81
Babble	34.94	37.41	39.9	42.62	13.47	14.97	16.47	17.65	2.33	3.2	4.22	4.63	0.76	0.81	0.81	0.94	16.61	17.27	20.41	21.19	3.5	3.98	4.7	4.89
Airport	39.06	40.91	42.09	43.39	13.72	15.69	17.2	17.47	2.33	2.79	4.09	4.49	0.78	0.77	0.89	0.95	16.15	17.11	20.07	21.64	3.53	3.67	4.51	4.75
Jet Plane	36.14	40.97	43.16	46.4	14.36	15.86	17.46	17.58	2.28	3.12	3.99	4.48	0.72	0.82	0.86	0.92	16.46	17.38	20.31	21.54	3.49	3.92	4.49	4.67
Street	37.55	40.98	41.72	42.91	13.68	14.67	15.66	17.53	2.29	2.98	4.23	4.58	0.8	0.86	0.89	0.94	16.24	17.18	20.1	21.1	3.5	3.89	4.51	4.74
Train Whistle	35.69	41.52	42.56	44.75	12.79	12.89	14.95	17.44	2.31	2.99	3.87	4.63	0.83	0.88	0.89	0.92	15.94	16.92	19.85	21.12	3.47	3.92	4.52	4.61
Restaurant	38.14	40.62	41.82	42.62	12.39	13.69	16.48	17.19	2.5	3.24	3.92	4.42	0.78	0.84	0.89	0.93	16.18	16.87	19.82	20.55	3.54	3.91	4.53	4.75
Car	33.015	38.19	40.86	41.63	11.08	12.36	14.69	16.97	1.98	2.97	3.61	4.46	0.67	0.79	0.82	0.9	12.76	15.68	18.04	19.85	3.36	3.56	4.36	4.5
Subway	33.985	37.33	39.34	41.48	11.63	12.81	15.34	17.37	2.08	2.98	3.72	4.51	0.69	0.8	0.84	0.92	12.95	15.93	18.82	19.95	3.24	3.68	4.41	4.53

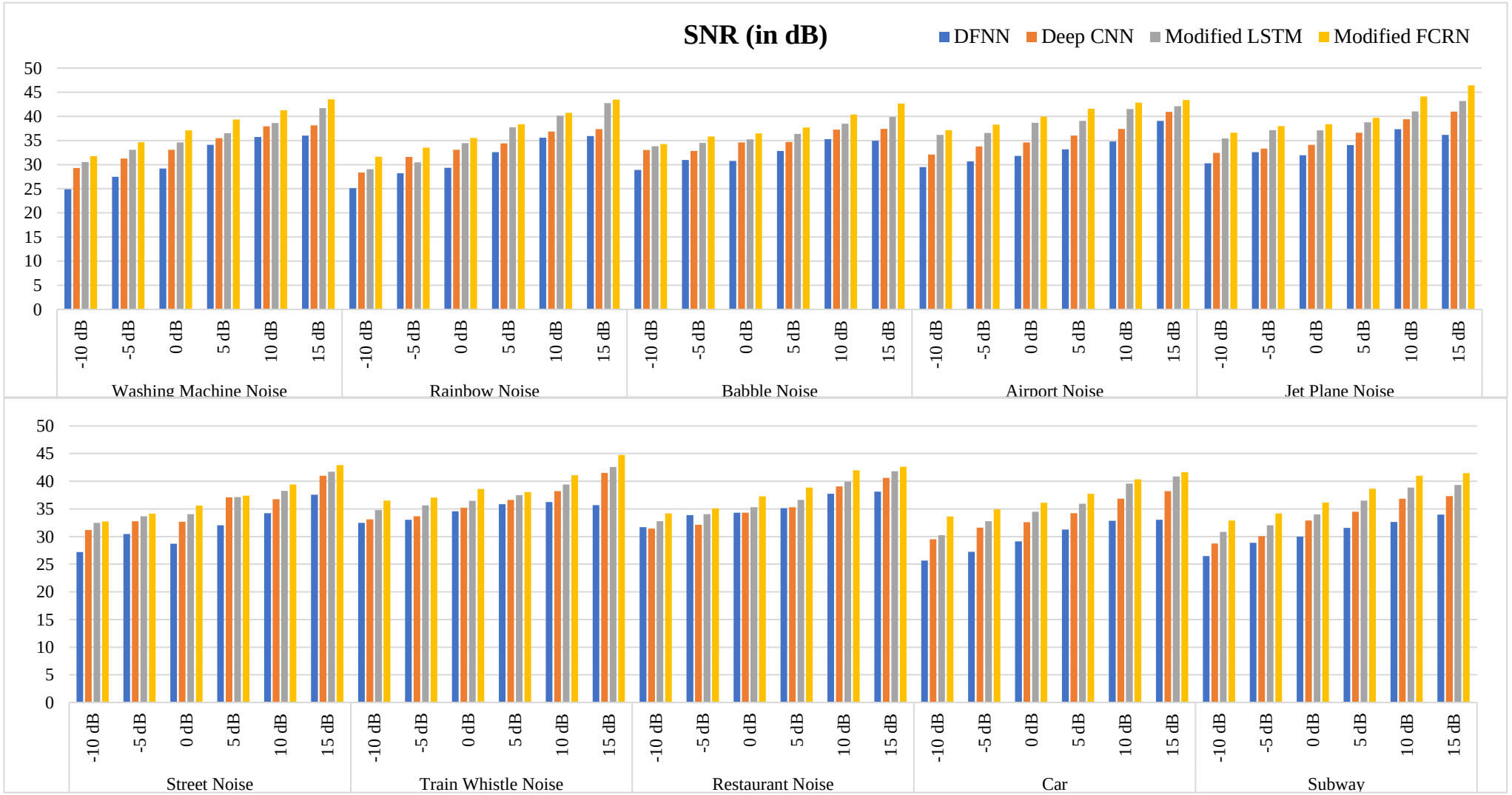


Figure 6.2 Comparative Analysis of SNR of DNN Algorithms for Alaryngeal Speech at Various Noise Types and Levels

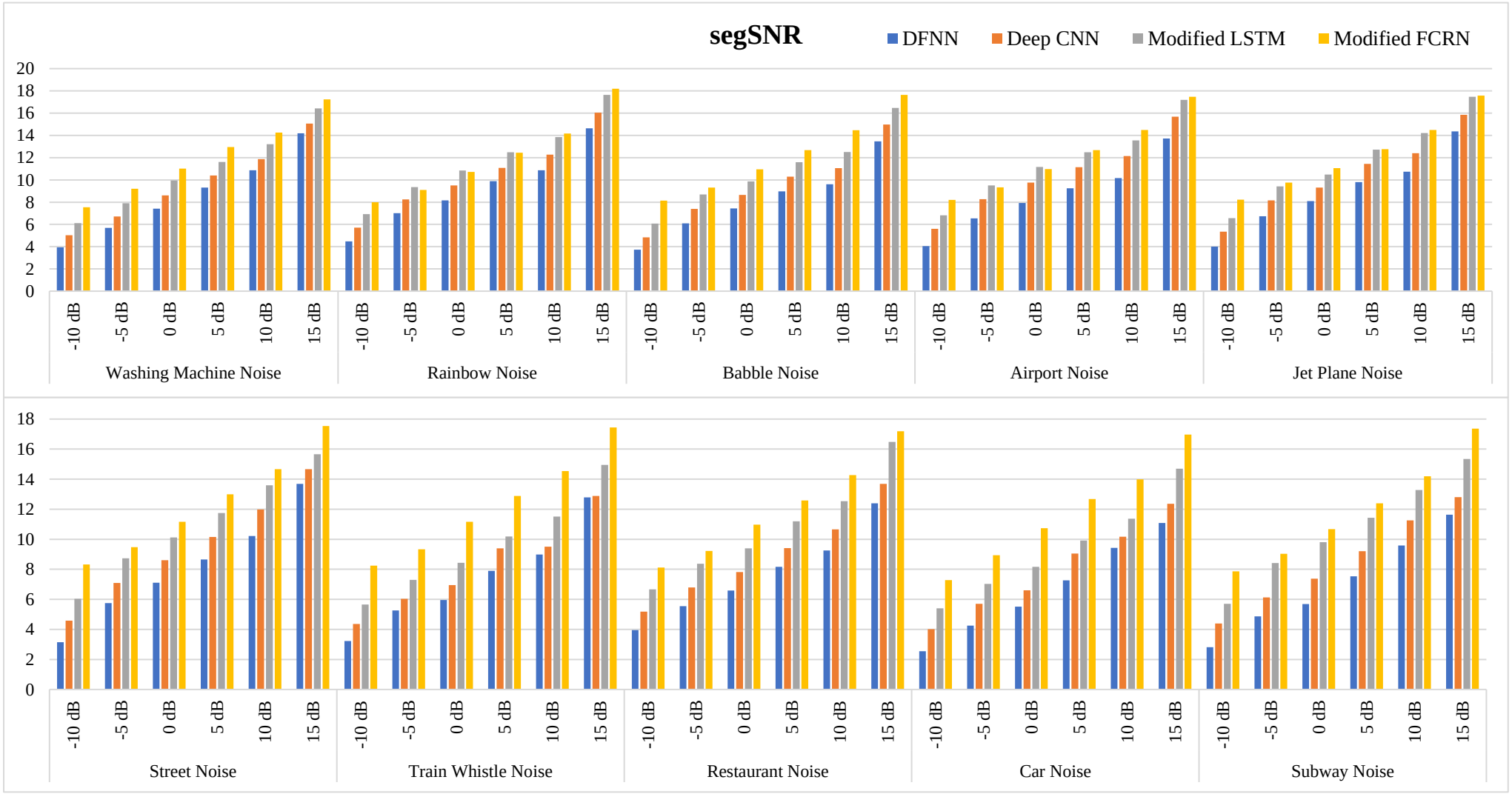


Figure 6.3 Comparative Analysis of segSNR of DNN Algorithms for Alaryngeal Speech at Various Noise Types and Levels

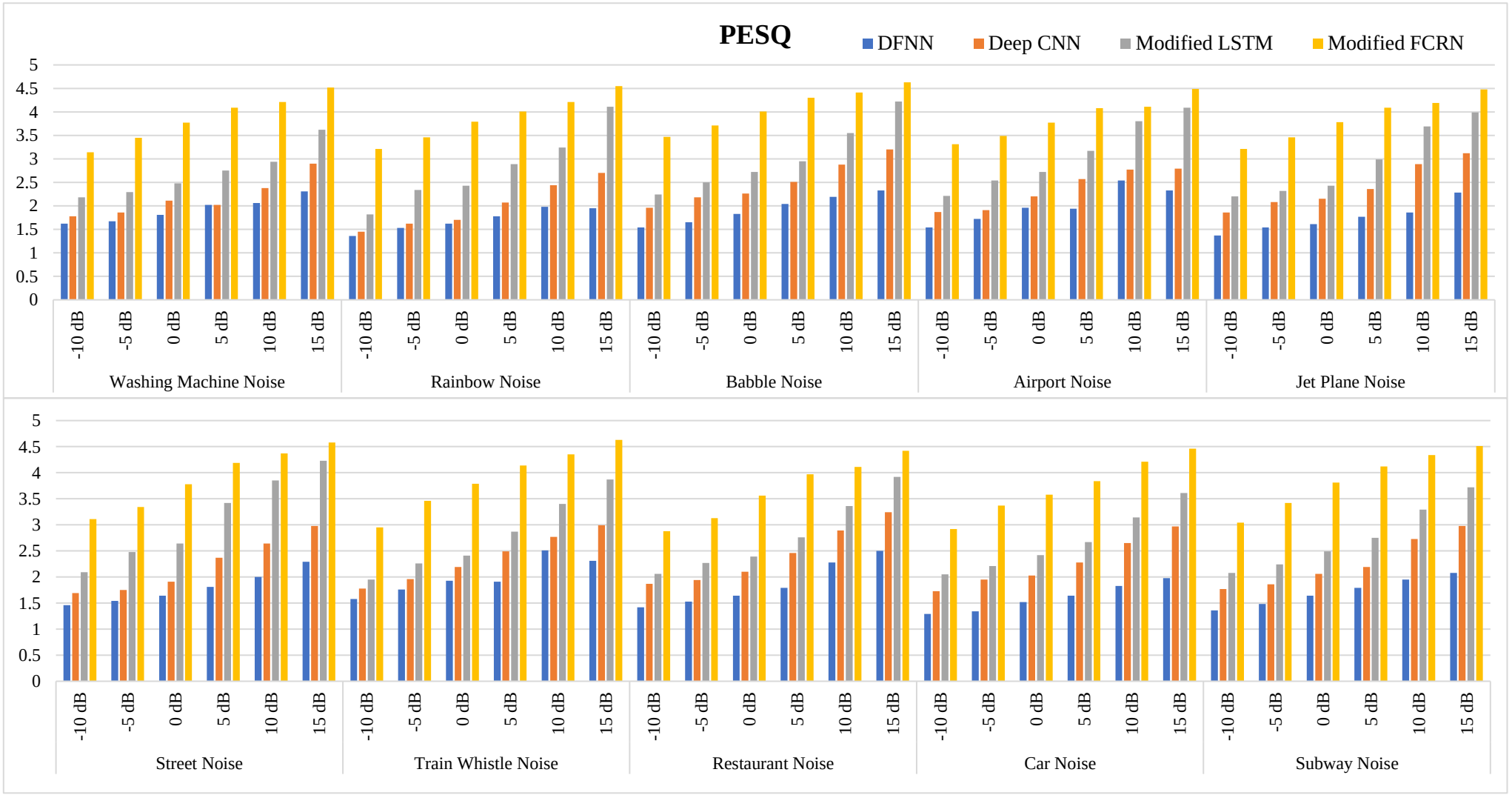


Figure 6.4 Comparative Analysis of PESQ of DNN Algorithms for Alaryngeal Speech at Various Noise Types and Levels

Figure 6.5 Comparative Analysis of STOI of DNN Algorithms for Alaryngeal Speech at Various Noise Types and Levels

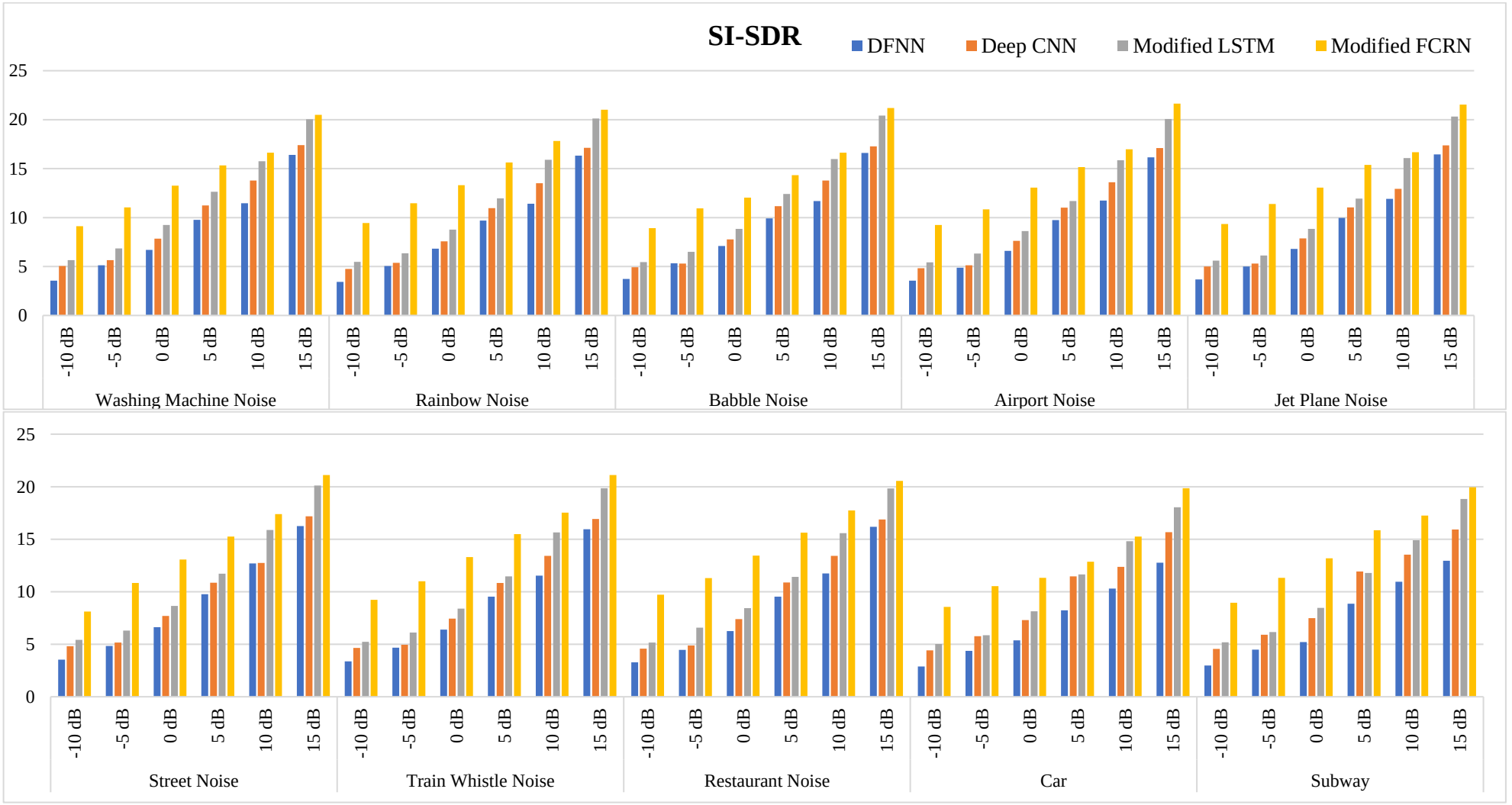


Figure 6.6 Comparative Analysis of SI-SDR of DNN Algorithms for Alaryngeal Speech at Various Noise Types and Levels

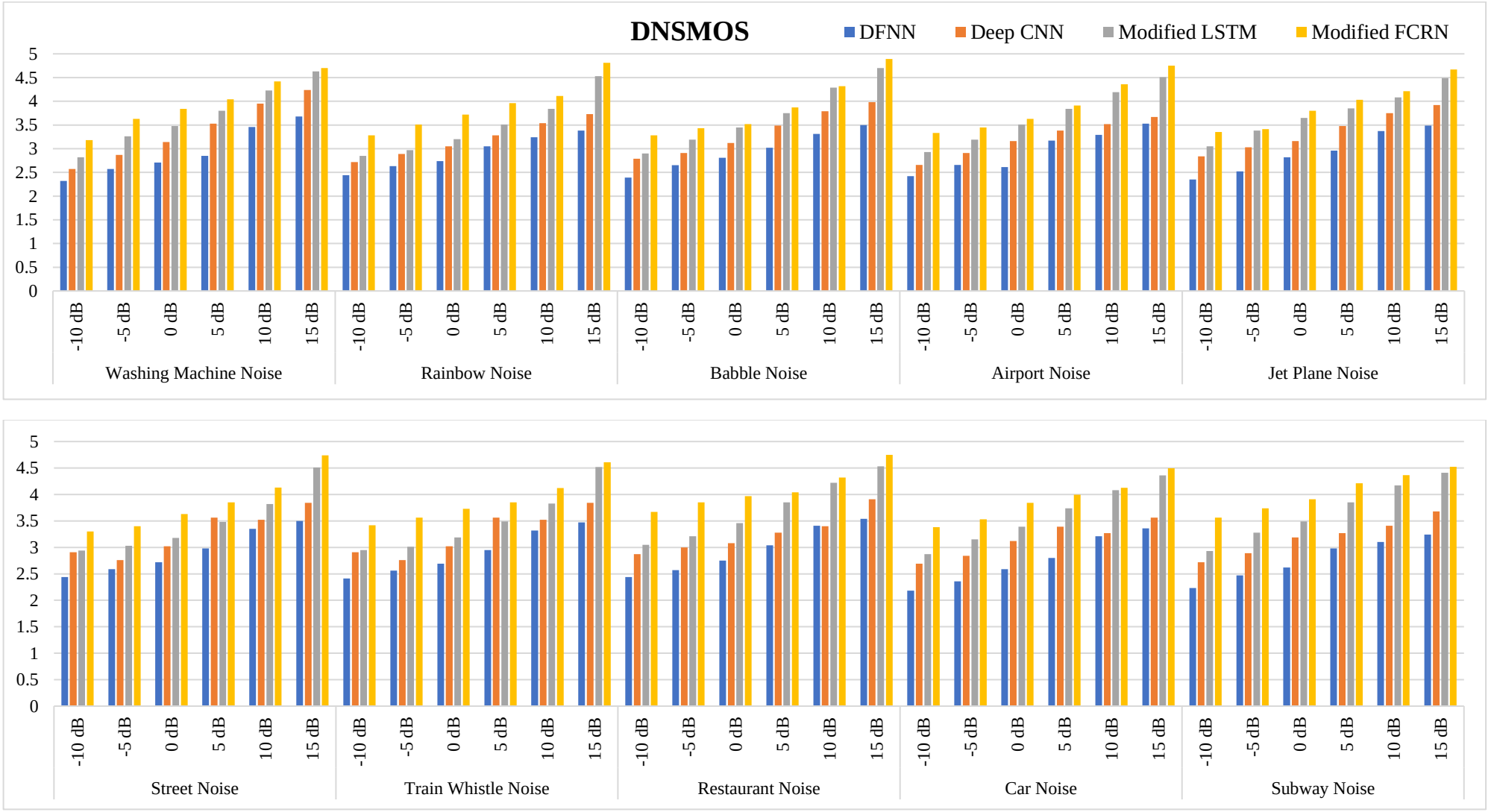
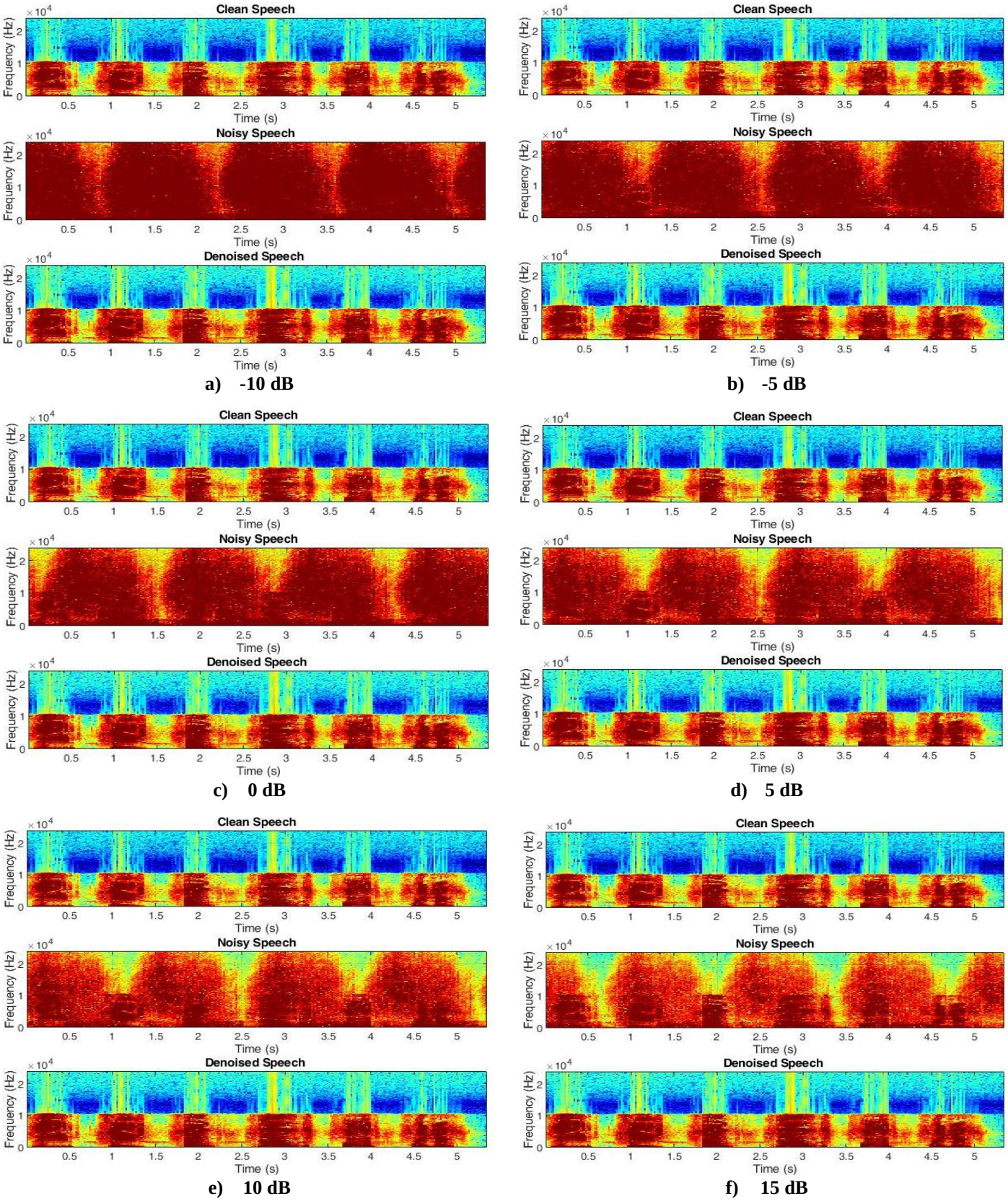


Figure 6.7 Comparative Analysis of DNSMOS of DNN Algorithms for Alaryngeal Speech at Various Noise Types and Levels

The spectrum of frequencies for Alaryngeal speech is visualized by the spectrograms shown in Figures 6.8 to 6.10 for various algorithms in different noises and noise levels. When comparing the noisy and denoised speech spectrograms, it is inferred that the noisy speech has high intensity represented with red color, and the denoised speech spectrogram has a mix of blue, yellow, and red that indicates the denoised speech and pauses in between the speech. The denoised speech spectrogram closely resembles the characteristics of the clean speech spectrogram. Regions where the noise is dominant in the noisy spectrogram show significant improvement in the denoised spectrogram based on the denoising ability of the deep learning algorithms. The denoised spectrogram indicates enhanced frequency components, reduced noise interference and improved temporal patterns compared to the noisy speech spectrogram.

Figure 6.8 represents the spectrogram image of jet plane noise at various noise levels of modified FCRN. It is characterized by a broad, intense energy distribution that spans across a wide frequency range, making it a particularly challenging type of noise. The jet plane noise has a high and wide energy distribution, often masking any underlying speech signal across most of the frequency spectrum. The noise is especially prominent in the lower to mid-frequency bands, which are crucial for speech intelligibility, making it difficult to distinguish the speech components. In an enhanced speech spectrogram, chaotic and continuous jet noise is significantly reduced, resulting in cleaner and sharper speech frequency bands. Figure 6.9 represents the spectrogram image of street noise for modified FCRN technique. The street noise spectrogram typically exhibits a complex and fluctuating energy pattern due to the diverse sources of sound found in a typical urban environment. These sources might include passing vehicles, honking horns, footsteps, and ambient chatter, all contributing to a varied and dynamic noise profile. Due to continuous background noise, street noise has a scattered energy distribution with peaks corresponding to loud, isolated sounds and more consistent energy spread in lower frequencies. This distribution can obscure speech signals, particularly in the lower frequency bands where much of the energy from vehicles and general street ambience resides. In enhanced speech, the dispersed patterns of street noise are significantly reduced, and the algorithm effectively filters out high-energy and sudden noises. The spectrogram of the enhanced speech shows sharp, well-defined bands of speech frequencies with reduced interference. Figure 6.10 represents the spectrogram image of

car noise for the modified FCRN technique. The car noise spectrogram is typically characterized by a steady, continuous energy pattern that reflects the constant hum of an engine, the sound of tires on the road, and occasional sharp bursts from car horns or sudden accelerations. This type of noise is pervasive and often masks essential elements of a speech signal. In the enhanced speech, the steady, low-frequency noise from the car is significantly reduced, as the modified FCRN algorithm effectively filters out the constant engine hum and the sporadic loud noises such as honking. The spectrograms obtained for all the four algorithms are given in Appendix 5 (Figures 1 to 40).



a) -10 dB b) -5 dB

c) 0 dB d) 5 dB

e) 10 dB f) 15 dB

Figure 6.8 Modified FCRN – Spectrogram Images of Jet plane Noise for Alaryngeal Speech at various Noise Levels

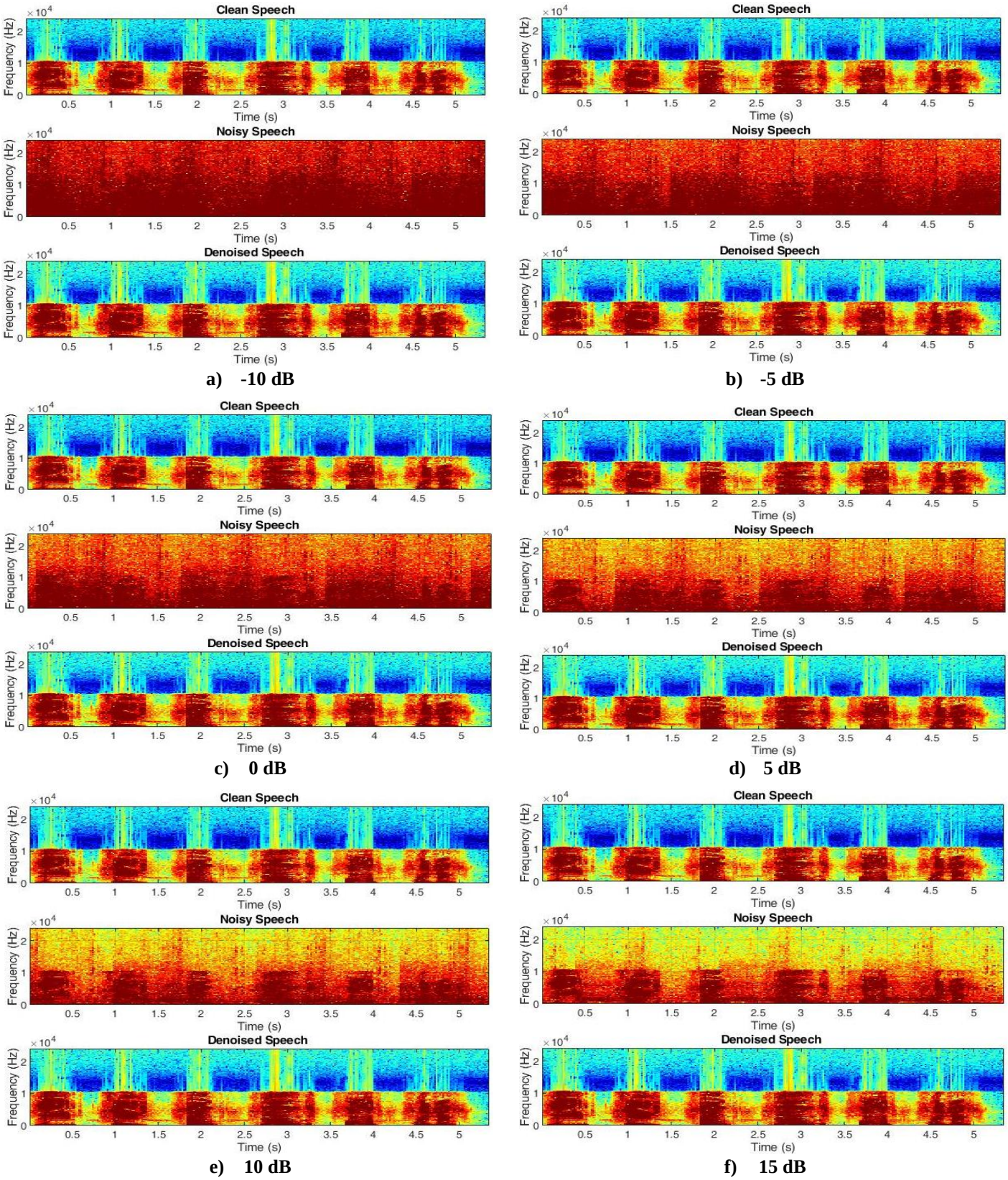
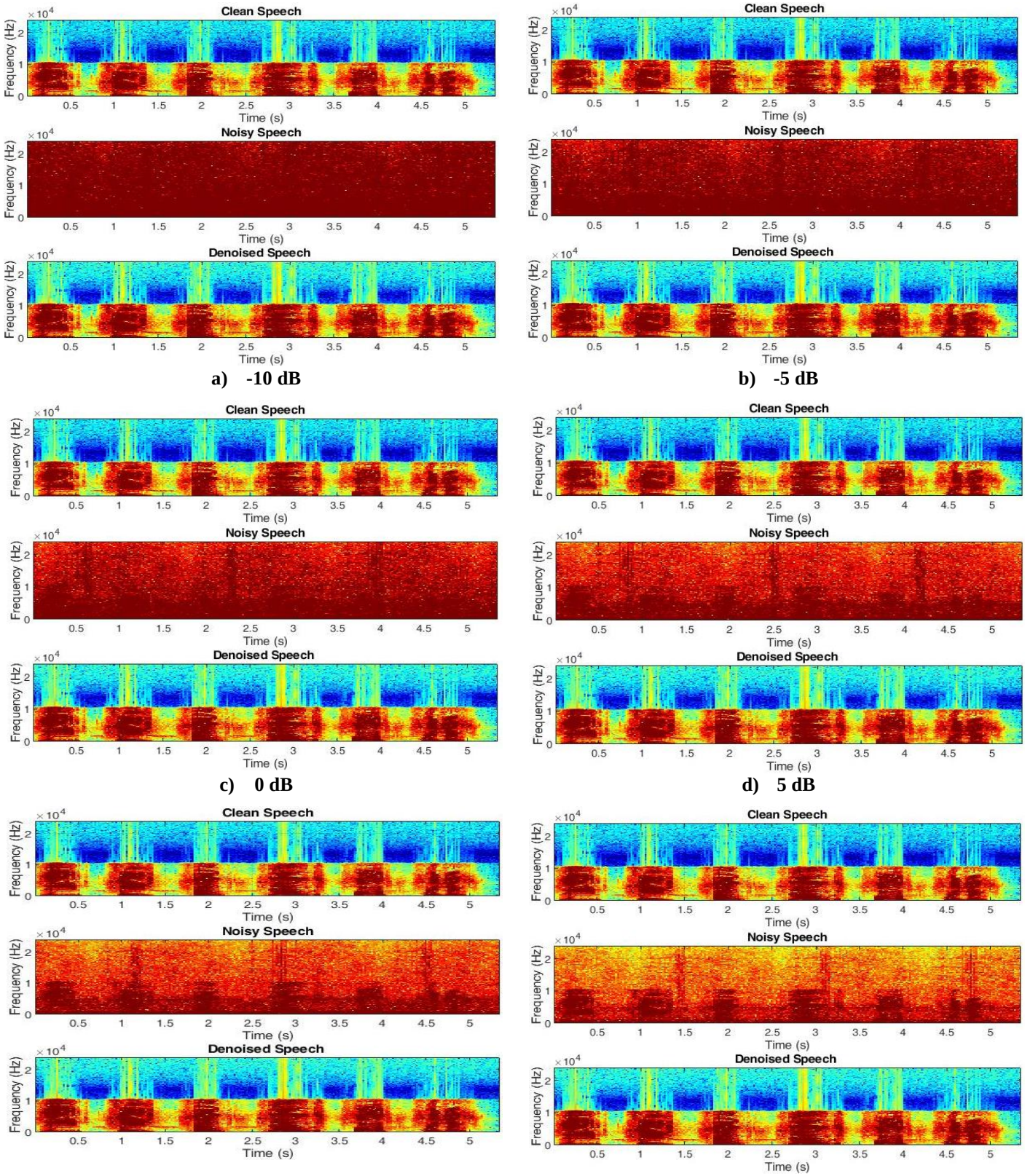


Figure 6.9 Modified FCRN – Spectrogram Images of Street Noise for Alaryngeal Speech at various Noise Levels



a) -10 dB b) -5 dB

c) 0 dB d) 5 dB

e) 10 dB f) 15 dB

Figure 6.10 Modified FCRN – Spectrogram Images of Car Noise for Alaryngeal Speech at various Noise Levels

6.7 WORD ERROR RATE

In order to validate the performance of the speech enhancement methods, the Harvard speech sentences spoken by laryngectomy patients are analyzed by calculating the Word Error Rate (WER) using Google Speech-to-Text API. This evaluation involved enhancing alaryngeal speech sentences using different deep learning algorithms. For each speaker i (i=1,2,…,10), the enhanced speech S_i was fed into the API, which converts the speech into textual data T_i. The ground truth G_i is the text of phonetically balanced sentences taken from Harvard speech sentences. The performance of the speech-to-text conversion was assessed by comparing the generated text T_i against the ground truth text G_i using the Word Error Rate (WER) as the metric. The WER is calculated using the formula:

$$WER = \frac{S+D+I}{N} \tag{6.1}$$

where:

S is the number of substitutions

D is the number of deletions

I is the number of insertions

N is the total number of words in the reference (ground truth) text G_i.

The WER is 100% when the alaryngeal speech is subjected to noise and improves when speech enhancement algorithms are applied. The speech sentences such as “Read verse out loud for pleasure” and “The stray cat gave birth to kittens” were taken to analyze WER. The total number of words spoken by each speaker is 13 words. The details of WER for each algorithm are given in Table 6.7. For instance, when 5 errors occur, it results in 3 substitutions, 1 deletion, and 1 insertion.

In the modified sentence, three substitutions have been made: “pleasure” has been replaced with “pressure”, “cat” has been changed to “cut”, “to” has been changed to “two”, the word “for” has been removed, and the word “the” has been inserted. The modified sentence is “Read verse out loud pressure” and “The stray cat gave birth to the kittens.”

Table 6.7 Word Error Rate

Speaker	DFNN		Deep CNN		Modified LSTM		Modified FCRN	
	Number of Errors	WER (%)	Number of Errors	WER (%)	Number of Errors	WER (%)	Number of Errors	WER (%)
1	9	69.23	5	38.46	2	15.38	1	7.69
2	9	69.23	6	46.15	2	15.38	3	23.08
3	8	61.54	4	38.46	2	15.38	1	7.69
4	8	61.54	7	53.85	3	23.08	2	15.38
5	9	69.23	5	38.46	2	15.38	1	7.69
6	8	61.54	6	46.15	3	23.08	3	23.08
7	9	69.23	7	53.85	4	30.77	3	23.08
8	8	61.54	4	30.77	2	15.38	2	15.38
9	8	61.54	5	38.46	2	15.38	2	15.38
10	8	61.54	5	38.46	2	15.38	1	7.69
Average		**64.61**		**42.30**		**18.45**		**14.61**

The WER for the speech signal enhanced by various algorithms is analyzed and shown in Figure 6.11. It is apparent that the alaryngeal speech signal enhanced by the modified FCRN algorithm results in a lower word error rate and increases the chances of recognition of speech sentences. The word error rate is 15%, and the reason behind this is the difficulty laryngectomy patients face in pronouncing sentences with the voice prosthesis device.

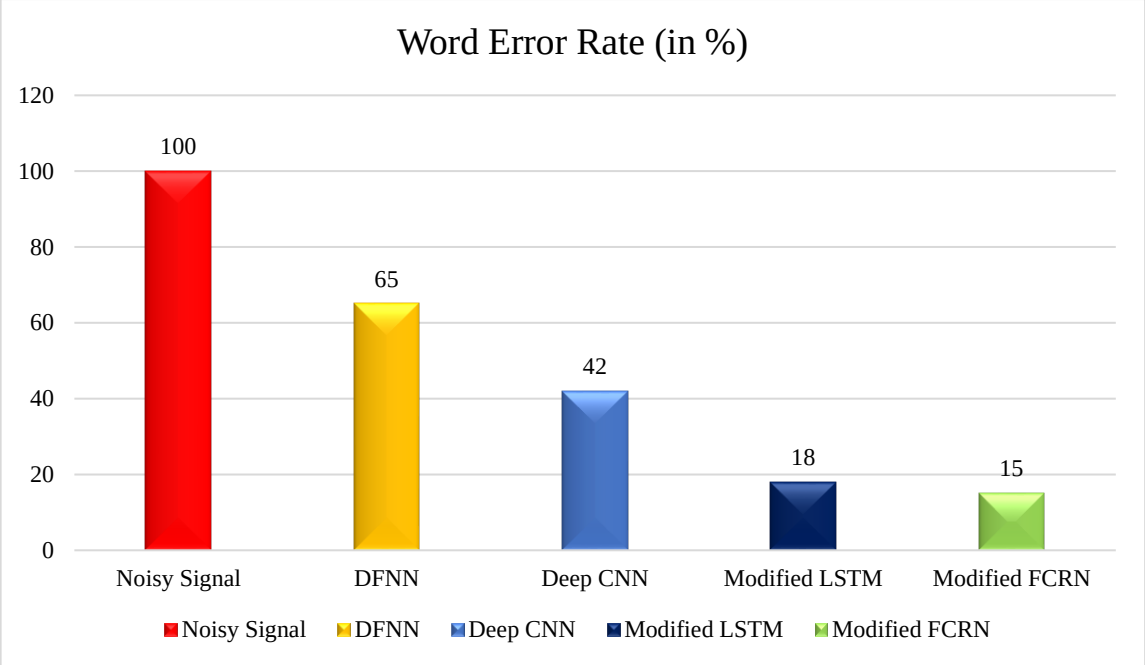


Figure 6.11 Validation of Speech Enhancement by Word Error Rate

6.8 MATLAB APPLICATION FOR ALARYNGEAL SPEECH ENHANCEMENT

The MATLAB App GUI for enhancing Alaryngeal Speech is shown in Fig 6.12. The app features a graphical user interface (GUI) that quickly loads alaryngeal speech recordings and selects noise types and levels. It is a user-friendly tool designed to enhance alaryngeal speech recordings in the presence of various types and levels of background noise. This app aims to improve the intelligibility and quality of alaryngeal speech signals by applying advanced speech enhancement techniques. The app helps to simulate the alaryngeal speech under different noise environments such as washing machine noise, airport noise, babble noise, street noise) at varying noise levels (-10dB to 15dB), and it provides enhanced speech as audio output. As alaryngeal speech is difficult to understand, especially in noisy environments, this app aims to address this issue by providing a customizable solution for enhancing alaryngeal speech recordings contaminated by real-world noise. The app employs the modified FCRN algorithm to enhance alaryngeal speech in noisy environments.

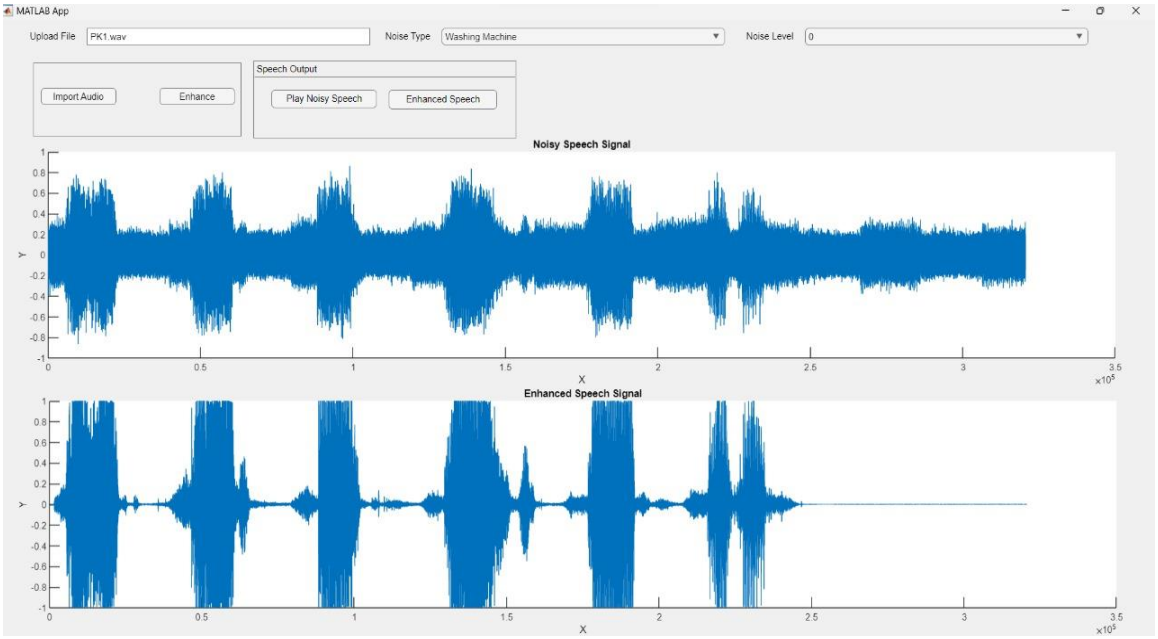


Figure 6.12 MATLAB App GUI for Alaryngeal Speech Enhancement

6.9 PARALINGUISTIC FEATURES

Paralinguistic analysis of voice is a crucial aspect of understanding human communication as it provides valuable insights into the emotional, social, and psychological aspects of spoken language. The paralinguistic features considered for analysis of speech are Pitch (Hz), Jitter (%), Shimmer (dB), Intensity (dB), Bandwidth, Formants and Harmonics to Noise Ratio (HNR). The features obtained in the analysis help compare the values of alaryngeal speech with normophonic speech.

6.9.1 Pitch

The pitch of speech is a fundamental and significant acoustic feature that plays a crucial role in interpreting and understanding spoken language. It refers to the highness or lowness of a person's voice during speech, and the frequency of the vocal cord vibrations determines it. Each individual has a unique pitch range, which is the span of frequencies they can produce during speech Resch et al., (2007). The average speaking voice commonly varies in pitch Sharifzadeh et al., (2010) from person to person, with a typical lower range of 60 to 160 Hz and a higher range of 160 to 320 Hz; for male speakers, frequency ranges between 60 to180 Hz and for female speakers 160 - 300 Hz (Re et al., 2012). Laryngectomy patients can achieve a more natural and varied pitch range with a voice prosthesis than esophageal speech. Adjusting airflow through the prosthesis allows

for better pitch control, leading to a greater pitch variation and the potential for more expressive speech. While Tracheoesophageal Prosthesis (TEP) provides more pitch variation than esophageal speech, it is essential to consider that the pitch range of laryngectomy patients may still differ from typical, pre-surgery pitch ranges. However, with proper training and rehabilitation, many patients can achieve a pitch range for effective communication.

6.9.2 Intensity

Intensity in speech refers to the loudness or volume of the sound produced during communication. It plays a crucial role in conveying emotions, emphasizing important information, and maintaining effective communication. The intensity of normophonic speech varies naturally based on several factors, including emotions, stress, the urgency of the message, and the distance between the speaker and the listener. The intensity of alaryngeal speech can vary based on the size and effectiveness of the voice prosthesis.

6.9.3 Jitter

Jitter is known as involuntary laryngeal behavior during phonation, which depends on the speaker's physical-psychological state and the linguistic structure of the language (Shahnaz et al., 2006). Jitter is a measure of the variation in the timing of vocal fold vibrations during speech. It is an acoustic measure used to assess the stability of the vocal fold vibrations, which can provide valuable information about speech quality. Jitter measures the cycle-to-cycle variability in the fundamental frequency of the vocal fold vibrations. It is typically expressed as a percentage and is often perceived as vocal instability or hoarseness when present in excessive amounts. Jitter values are generally low, indicating a stable and consistent vibratory pattern. This contributes to a smooth and pleasant-sounding voice during speech. Jitter analysis is relevant in assessing the stability of the vibrations generated by the prosthesis. A well-functioning prosthesis should produce relatively stable vibrations, leading to smoother speech.

6.9.4 Shimmer

Shimmer measures the variation in the amplitude or intensity of vocal fold vibrations during speech. The quantification of shimmer provides a means to differentiate between pathological and normal voices, facilitating the early detection of glottal diseases

(Shahnaz et al., 2006). It is an acoustic parameter that assesses the stability of the voice by quantifying the cycle-to-cycle fluctuations in the amplitude of the vocal fold vibrations. Shimmer is relevant for analyzing both normophonic speech (speech produced by individuals with intact vocal folds) and alaryngeal speech (speech produced by individuals who have undergone laryngectomy). Shimmer measures the variation in the amplitude of the sound wave during each vocal fold vibration cycle, typically expressed as a percentage. Shimmer values are generally low for a normal person, indicating a stable and consistent amplitude of vocal fold vibrations. This contributes to a smooth and steady voice quality during speech. An increase in shimmer can be indicative of various voice disorders. When analyzing alaryngeal speech, shimmer helps assess the stability of the amplitude of vocal fold vibrations generated by the prosthesis.

6.9.5 Harmonics to Noise (HNR) Ratio

Harmonics-to-noise ratio (HNR) is an acoustic measure used to assess the balance between harmonic components and noise in the speech signal. Harmonics are the periodic vibrations of the vocal folds that create the fundamental frequency and its multiples (overtones) during speech production (J.-W. Lee et al., 2014). Noise, conversely, represents the non-periodic, aperiodic components of the speech signal. In normophonic speech, the vocal folds vibrate regularly, producing a series of harmonics that create the fundamental frequency and its overtone frequencies. These harmonics contribute to the periodic, tonal aspects of the speech signal.

On the other hand, noise in the speech signal can be caused by turbulence in the airflow, fricative sounds, or other non-tonal components. The HNR is typically high for a normal person, indicating a strong presence of harmonics in the speech signal compared to noise. This results in a clear and smooth voice quality during speech. For individuals who use a voice prosthesis to speak, HNR analysis can help assess the balance between harmonic components and noise in the speech signal generated by the prosthesis. A well-functioning prosthesis should produce a relatively high HNR, producing a clear and intelligible voice quality.

6.9.6 Formants

Formants are important acoustic features of speech that play a significant role in shaping the quality of the vowel sounds produced by the human vocal tract. They are resonant frequencies at which the vocal tract amplifies certain harmonics of the sound wave produced by the vocal cords. Formants are essential for speech perception and distinguishing between vowel sounds (Gowda et al., 2020).

When the vocal cords vibrate during speech production, they generate a complex sound wave that contains a fundamental frequency and its harmonics. As this sound wave travels through the vocal tract, the shape and length of the vocal tract cause certain harmonics to be amplified or dampened, leading to forming formants. Each formant corresponds to a specific resonant frequency of the vocal tract. Formants primarily contribute to the distinct quality of vowel sounds in speech. Different vowel sounds are characterized by varying formant frequencies.

Formant analysis of alaryngeal speech is a valuable tool for assessing the acoustic characteristics of speech produced by individuals who have undergone laryngectomy and lack a larynx and natural vocal folds. Formant analysis helps evaluate the resonance and articulatory characteristics of speech, thereby providing insights into the quality and intelligibility of the alaryngeal voice.

6.9.7 Analysis of Paralinguistic Features

Table 6.8 compares the speech sentence “Read Verse Out Loud for Pleasure” spoken by a normal person and someone who has undergone a laryngectomy. Based on the analysis of the values of paralinguistic features, alaryngeal speech tends to have higher jitter and shimmer values than normal speech, representing the irregularities in vocal fold vibrations. This is due to the altered vocal production mechanism in alaryngeal speakers. The pitch values in alaryngeal speech are generally higher than those in Normal Speech, while the intensity values are also higher. This could indicate altered vocal tract configurations and increased effort in alaryngeal speech production. The lower HNR values of alaryngeal speech than normal speech indicate a higher noise level or a less clean

speech signal. This is due to lacking a vocal fold source in alaryngeal speech. Formant frequencies (F1, F2, F3, and F4) also differ between the two speech categories. These differences are related to the changes in vocal tract resonances and articulatory movements in alaryngeal speakers. The frequency bands (B1, B2, B3, and B4) also exhibit variations between the two speech categories, which may indicate differences in the spectral characteristics of the speech signals. Based on the paralinguistic features that exhibit variation in alaryngeal speech, re-engineering the prosthetic device design is possible, making it sound similar to normophonic speech.

Table 6.8 Comparison of Paralinguistic features of Normophonic and Alaryngeal Speech

Words of the Sentence	Speech Category	Paralinguistic Features												
		Jitter (%)	Shimmer (dB)	Pitch (Hz)	Intensity (dB)	HNR (dB)	F1	F2	F3	F4	B1	B2	B3	B4
READ	Normophonic	1.01	1.10	137.5	64.84	9.91	372.71	1710.35	2090.61	2736.59	74.31	570.34	109.34	142.23
	Alaryngeal	1.38	1.29	170.94	80.15	3.56	414.49	2794.98	3163.4	4189.42	145.84	203.63	138.07	292.54
VERSE	Normophonic	1.36	1.2	138.69	68.37	9.48	544.74	1327.93	1812.64	2322.53	80.13	54.94	54.66	398.15
	Alaryngeal	1.52	1.64	147.66	76.28	2.18	558.06	2514.58	3086.32	4099.18	249.58	183.87	183.14	260.24
OUT	Normophonic	1.2	0.82	133.10	65 .36	8.99	731.26	1393.88	1842.64	2523.58	73.74	144.83	322.28	96.03
	Alaryngeal	1.24	0.99	135.46	78.82	2.96	850.02	1569.9	2941.71	4153.49	275.25	130.54	125.14	164.98
LOUD	Normophonic	1.29	1.25	112.93	64.37	7.51	615.70	1255.95	1908.41	2518.37	88.79	81.02	557.75	55.03
	Alaryngeal	1.6	0.86	151.92	80.79	4.77	777.08	1641.52	3022.22	4281.38	128.4	141.6	63.78	487.03
FOR	Normophonic	1.17	1.62	101.47	71 .64	8.63	716.74	1364.46	1792.28	2782.53	81.62	148.41	317.62	127.72
	Alaryngeal	1.61	0.83	135.19	76.90	5.54	863.45	1293.63	2828.94	4312.88	376.4	293.61	216.25	495.39
PLEASURE	Normophonic	1.23	1.32	116.35	62.82	8.24	523.52	1271.25	1656.82	2651.88	58.07	212.3	140.72	63.55
	Alaryngeal	1.65	1.51	119.88	77.02	1.37	805.54	2056.8	2949.63	4392.24	1205.04	723.99	237.71	372.03

6.10 SUMMARY

The same approach has been used with alaryngeal speech as that of speech data from the database. Among the four algorithms (DFNN, Deep CNN, modified LSTM, and modified FCRN), the deep learning architecture with modified FCRN performed better than the other algorithms. Hence, it is again proved that data-driven approaches outperform simple statistics-driven approaches.

By observing the WER of the enhanced alaryngeal speech, the ability to understand the speech is validated. MATLAB App's design is a user-friendly tool to enhance alaryngeal speech recordings. The paralinguistic features analyzed will help re-engineer the design of the voice prosthesis.