SNR-Selection-Based-Data Augmentation for Dysarthric Speech Recognition

Sarkhell Sirwan NAWROLY[1]*, Decebal Gheorghe POPESCU[1], Mariya Celin THEKEKARA ANTONY[2], Actlin Jeeva MUTHU PHILOMINAL[3]

[1] Faculty of Automatic Control and Computer Science, University Politehnica of Bucharest, 060042, Romania
sarkhell.sirwan@spu.edu.iq (**Corresponding author*), decebal.popescu@cs.pub.ro

[2] Department of Electronics and Communication Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, 600062, Tamil Nadu, India
mariyacelinta@veltech.edu.in

[3] Department of Electronics and Communication Engineering, Sri Sivasubramaniya Nadar College of Engineering, Kelambakkam, 603110, Tamil Nadu, India
actlinjeevamp@ssn.edu.in

Abstract: With the recent advances in Automatic Speech Recognition systems, the lifestyle of normal people has become more convenient. However, for a population like the speech disordered community, the efficiency or the use of such ASR systems is very limited because these ASR systems are not trained or modelled with speech data pertaining to medically impaired people. The difficulty in training such ASR systems lies in the poor availability of data. To handle this issue, an approach like data augmentation for dysarthric speech recognition was analyzed in this paper. Noise is a source that is freely available in abundance. In speech recognition, noise has been used for developing a robust ASR system. This paper focuses on using noise as a source for data augmentation for increasing the number of dysarthric speech samples and improving the performance of speech recognition systems. The core idea behind this research work is that when a sound is combined or enhanced with another sound, its impact is noticeable only if both sounds have the same frequency range. Therefore, understanding the characteristics of each noise sample and adding them appropriately to the dysarthric speech data to create new samples of dysarthric speech data is the proposed method for increasing the number of dysarthric speech examples. Initially, noise samples were selected that do not affect the dysarthric speech frequency range. At a particular signal-to-noise ratio (SNR) the noise-augmented dysarthric speech examples were then used for training dysarthric speech recognition systems by employing hybrid DNN-HMM-based systems for isolated dysarthric speech examples. After noise selection-based data augmentation, it was observed that the word error rate (WER) was reduced by 7% for all the categories of dysarthric speakers in comparison with the WER for the ASR system trained without data augmentation. Since this approach used low-frequency noises as a source for data augmentation, the number of augmented examples was not restricted to a limit; the higher the number of low-frequency noises within a selective SNR range, the better the augmented examples. Further on, this approach used the selected dysarthric speech examples for augmentation, making the augmented examples not lose the dysarthric speakers' identities.

Keywords: Data augmentation, Noise selection, SNR-based, Dysarthria, Speech recognition.

1. Introduction

Dysarthria is a speech disorder caused due to a range of motor control conditions such as cerebral palsy, amyotrophic lateral sclerosis, stroke and traumatic brain injuries (Darley et al., 1975). The speech of dysarthric people eventually becomes slurred or unintelligible due to the uncontrolled and uncoordinated articulatory movements. To handle their speech disability, augmentative and alternative communication (AAC) aids were developed to support their communication needs with regard to their peers. Amongst these, speech assistive aids developed using automatic speech recognition (ASR) systems would be of great support to them by making their life easy and comfortable. However, ASR systems developed today are not meant for them because they have not been trained with the speech data of people with speech disorders. To train such an ASR system specific to these speakers, speech data from these speakers is highly required. The rule of thumb according to which the more speech data, the greater the performance applies to the training of dysarthric ASR systems as well. On that note, gathering such a huge corpus for these speakers is difficult. There are few corpora available on dysarthric speech data; however, they are limited. An abundant corpus, as it is the case for normal speakers, is not available for dysarthric speakers. This is because the collection of dysarthric speech data is not as straightforward as for normal speakers since it involves recording speech data from speech disordered speakers who are physically and intellectually disabled. Eventually, this task of collecting data from the disordered speakers becomes tiring and difficult.

Several dysarthric speech corpora available for research on dysarthric speech are free or they can be provided for a fee by the Linguistic Data Consortium. The Nemours dysarthric speech corpus (Menendez-Pidal et al., 1996) consists of dysarthric speech data for only 11 males who have uttered only 74 sentences each of six words in length; they make up less than 3 hours of speech data. Comparatively, a bigger database is the TORGO dysarthric speech

https://doi.org/10.24846/v32i4y202312

corpus (Rudzicz, Namasivayam & Wolff, 2012), which has dysarthric speech data with a duration close to 15 hours. Currently, the largest dysarthric speech corpus that is available for free is the UA dysarthric speech corpus (Kim et al., 2008), that contains 102.7 hours of speech data recorded from 29 speakers, of which 16 are dysarthric speakers, while the remaining 13 are healthy speakers. All the above discussed dysarthric speech corpora are in the English language. The SSN-TDSC (Thekekara Antony et al., 2016) is a dysarthric speech corpus that includes dysarthric speech data collected from 20 dysarthric speakers in the Tamil language of India. This corpus is mainly designed for developing assistive aids for dysarthric speakers with data from close to 25 hours of dysarthric speech. All the above discussed dysarthric speech corpora consist of nearly 50 to 60 hours of dysarthric speech data, which is comparatively far from the speech data required to train successful ASR systems such as Siri, Alexa or Google Voice Assistant. This is because collecting speech data pertaining to medically impaired people is more challenging than when it is collected for normal speakers. However, as the general thumb rule says, for any successful ASR system to be developed as a product it requires a humongous amount of speech data for training. Data augmentation approaches were developed to handle these data insufficiencies in dysarthric speech data. This research work focuses on data augmentation for dysarthric speech recognition, and the following section discusses previous research on data augmentation for dysarthric speech data.

The rest of the paper is organized as follows. Section 2 discusses previous research on data augmentation for dysarthric speech data. Section 3 discusses the dysarthric speech corpus and noisy data used in this paper. Section 4 provides an analysis of noisy signal data. Section 5 explains the augmentation of noise onto dysarthric speech data at various dB levels. Section 6 discusses the training of a DNN-HMM-based ASR system for the augmented speech data. Section 7 compares the proposed system with the systems presented in the previous works related to data augmentation. Finally, Section 8 concludes this paper.

2. Related Research

There is limited research on data augmentation targeting disordered speech recognition. For dysarthric speech data, data augmentation is achieved by transforming speech data pertaining to healthy speakers to sound like speech data from dysarthric speakers or by performing minimal transformations with the available dysarthric speech data to obtain transformed versions of the dysarthric speech data. Using a normal speaker's speech data, (Salamon & Bello, 2017) augmentation is achieved by synthetically generating speech examples by altering the speech rate using different speed factors, time stretching, pitch shifting, dynamic range compression, combining the data with environmental background noise, etc. In (Vachhani, Bhat & Kopparapu, 2018) data augmentation using temporal and speed modifications in normal speech to produce synthetic speech matching the characteristics of dysarthric speech was proposed. The synthetically generated dysarthric speech was classified using a random forest classifier and trained with a hybrid DNN-HMM-based dysarthric speech recognition system. The WER reported for the ASR systems without data augmentation and with time stretching based data augmentation and tempo-based data augmentation is 29.06%, 26% and 26.15%, respectively. Apart from speed perturbation in (Jiao et al., 2018), the speech data of the normal speaker is modified using a three-step conversion technique, which includes speaking rate modification, spectral feature transformation using deep convolutional generative adversarial networks (DC-GANs) and pitch modification. In this case, the classification accuracy improved by approximately 10% after data augmentation. Voice conversion is another approach (Jin et al., 2021) used for handling the data sparse conditions in dysarthric model training. However, the above discussed works focused on obtaining an equivalent amount of dysarthric speech data from a normal speaker's data were performed mostly for the mild and mild-to-moderate categories and not addressed to the remaining dysarthric speaker categories. This is because, when synthesizing speech from the normal speaker's speech data for the moderate and severe types of dysarthric speakers, there are chances that the dysarthric speaker's identity/ errors are not incurred in the transformed speech.

To address these difficulties, transformations using dysarthric speech data were performed to obtain new training data samples. In (Xiong et al., 2020), a 3-fold speed perturbation technique is used on the available dysarthric speech data to increase the size of the training data. The increased data size is trained using a factored form of time-delay neural networks (TDNN-F), and the average WER was about 30.76%. Additionally, virtual Microphone array synthesis and Multi-Resolution feature extraction (VM-MRFE)-based data augmentation (Thekekara Antony et al., 2020) is used to increase the training data size by directly transforming the dysarthric speech data. In this technique, initially using a single source example, virtual microphone array (MA) signals are synthesized, and with the synthesized MA signals, multi-resolution feature extraction (MRFE) is performed. From a single source signal, the data size is increased seven times, using virtual microphone array synthesis with an array size of 7. With multi-resolution feature extraction, the features of the augmented data are extracted with multiple window sizes varying from 12 ms to 20 ms in steps of 2 ms, yielding five different resolutions. Therefore, this two-level data augmentation technique increases each source example 35 (7*5) times. This task is carried out for both isolated words and continuous speech and proved to obtain the best results for continuous dysarthric (Thekekara Antony et al., 2023) speech augmentation over the traditional methods. However, it was discussed that the augmented data size cannot be increased further, as it reduces the signal's energy. Recently, a research paper concentrating on combining both approaches by using both normal speakers and dysarthric speakers to develop a data augmentation approach for dysarthric speakers has been published. A two-stage data augmentation (Bhat, Panda & Strik, 2022) has been developed where, in the first stage using Deep Autoencoder (DAE), various speech perturbations are performed on the speech data pertaining to healthy speakers and in the second stage using SpecAugment techniques modifications to the dysarthric speech data are performed. An absolute improvement of 16% in word error rate (WER) was achieved over a baseline with no augmentation.

These above listed approaches reduce the error rate by increasing the training data size and retaining the dysarthric speaker's speech identity. However, these approaches provide only a limited quantity of augmented samples, like 6 or 35, from a single training example. In contrast, huge samples are required to develop an effective ASR system that retains the speaker's identity.

Though data augmentation serves the purpose of handling data sparsity by performing transformation techniques, retaining the identity of the dysarthric speaker and still increasing enough data samples is challenging. Noisy data is one source already abundant in the literature (Muthu Philominal et al., 2020). Injecting noise to the available training examples and creating new samples of the source data is the core idea of this research work. However, by its nature, noisy data is disordered or corrupted data that cannot be understood and interpreted correctly. Injecting such data into the disordered speech data can make it even more corrupted and less intelligible. However, by understanding and analyzing the frequency range of each variety of noisy data, a better augmentation is still achievable. In (Borrie et al., 2017), the relationship between speech in noise and dysarthric speech data was analyzed, and it was found that they are positively associated or correlated. The noisy data used in this study was speech shaped noise (SSN) data, and the authors understood from a perceptual analysis that listeners who succeed in understanding speech in noise are the same listeners who successfully understand dysarthric speech. Hence, the conclusion was that adding SSN to normal speech data made it look like dysarthric speech data. As such, it is better to add noise to the dysarthric speech data for retaining the dysarthric speaker's identity and performing augmentation. However, using a noise type like SSN may not serve this purpose, as it adversely affects the dysarthric speech data that is already distorted. Anyhow, apart from SSN, there are other categories of noise data; this paper concentrates on identifying the suitable category of noise to use as a source of data augmentation. Identifying the noise source and augmenting it with the dysarthric speech data such that it doesn't lose the identity of the dysarthric speaker and obtaining a reduced word error rate (WER) is the main objective of this research paper.

3. Speech Corpus Used

The UA dysarthric speech corpus is used for the current work to experiment with the analysis of isolated word dysarthric speech data. The UA dysarthric speech database (Kim et al., 2008) includes dysarthric speech data uttered by 19 speakers with cerebral palsy and 13 normal speakers. The speakers with cerebral palsy are classified into four classes based on their speech intelligibility levels: high, moderate, low and very low. Each dysarthric speaker has uttered 765 isolated words recorded in a laboratory environment at a sampling rate of 48 kHz.

For the noisy data analysis, NOISEX-92 database (Varga & Steeneken, 1993) is used. Noise samples such as babble, factory, pink, benz, car, bus and volvo noises were chosen for the analysis. The noises were picked such that they cover a wide frequency range.

4. Analyzing Noise to Understand Its Characteristics for Performing Data Augmentation

The speech frequency range typically falls between 500 Hz to 4000 Hz. When a noise signal lies within this range, it can mask the speech frequency components in the speech data, leading to further corruption. Therefore, it is understood that the effect of noise frequency is minimal when it is added at the noise range beyond the specific speech frequency range.

LP-based formant extraction method is used, with a linear predictor of order 20. To analyze the frequency range of noise having the highest energy, histograms of the dominant frequency components are extracted and plotted for each noise category under consideration. Histograms for the noise frequencies like babble, factory, pink, benz, car and volvo were extracted and plotted in Figure 1. By observing Figure 1, it can be said that pink noise, babble noise, and factory noise have dominant frequency ranges of up to 2500Hz, which can affect or mask the speech frequency components in the speech signal upon augmenting. On the other hand, car noise also has a frequency range between 800 - 1000Hz. Still, its dominant frequency strength in this range is notably weaker, as depicted in Figure 2, compared to the dominant frequency strength of factory noise in Figure 8, pink noise in Figure 9 and Babble noise in Figure 10. Interestingly, volvo noise, benz noise, golf noise, and bus noises in Figures 3, 4, 5, 6, and 7, respectively, show very low energy in the high frequency range, indicating that the probability of masking speech components by these noises is minimal in comparison with other noise categories.

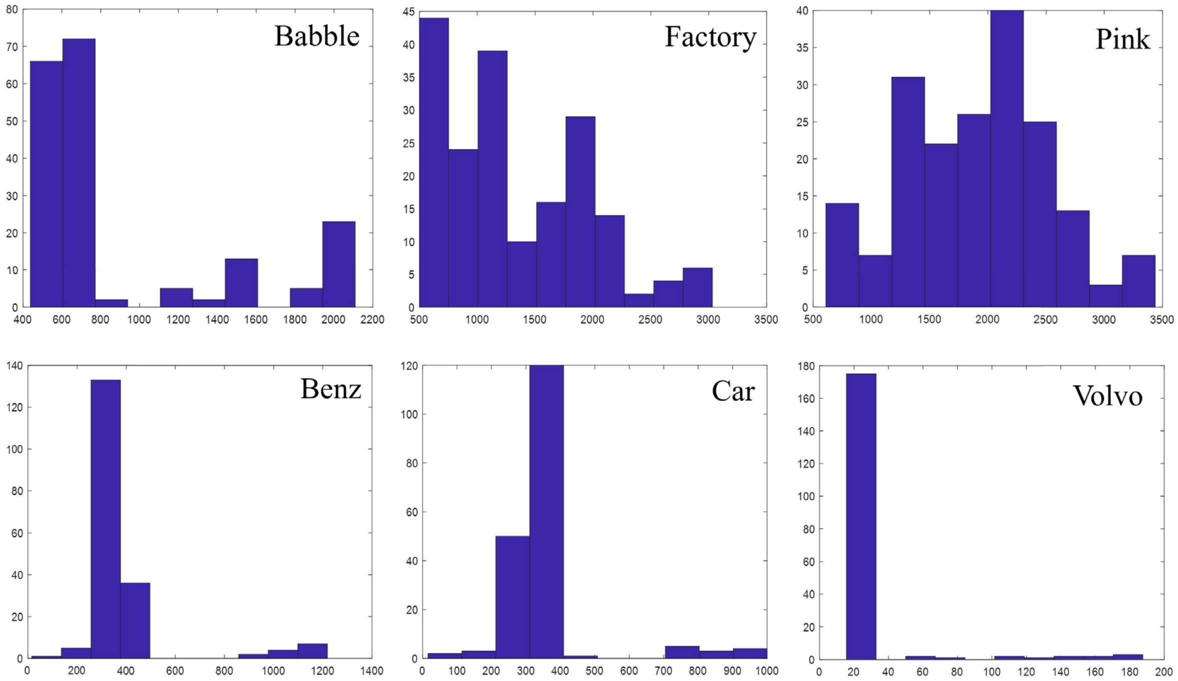

Figure 1. Histograms of the noise ranges from Noisex-92 showing the dominant frequency ranges

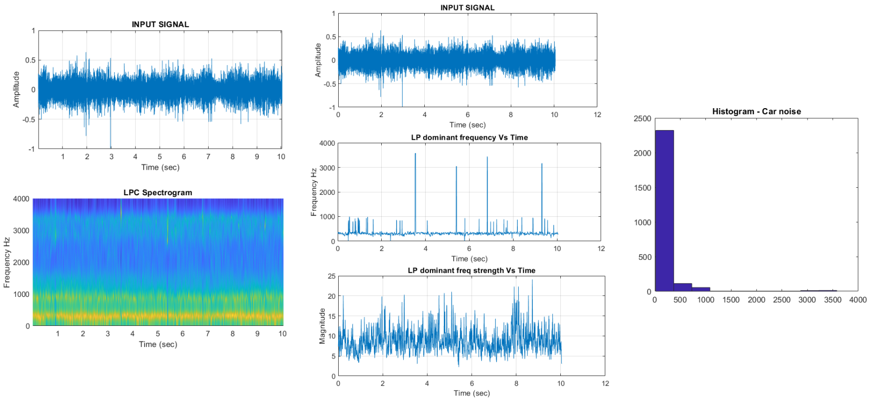

Figure 2. LPC spectrogram, LPC dominant frequency and strength of the dominant frequency for Car noise

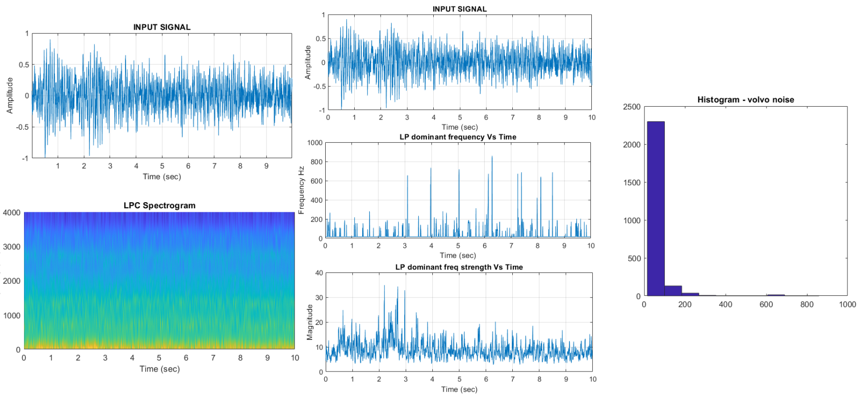

Figure 3. LPC spectrogram, LPC dominant frequency and strength of the dominant frequency for Volvo noise

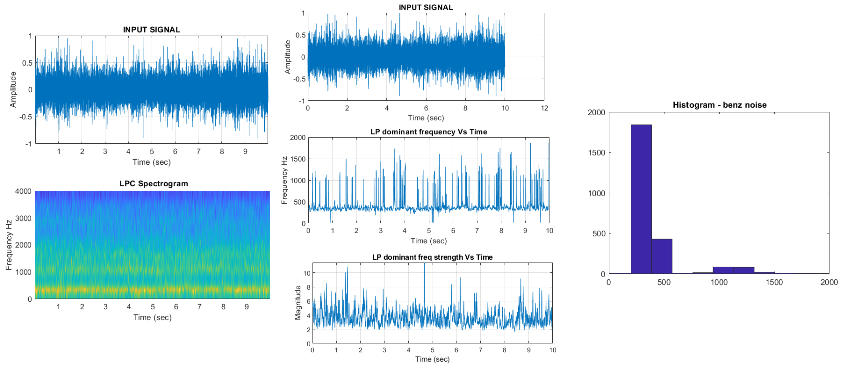

Figure 4. LPC spectrogram, LPC dominant frequency and strength of the dominant frequency for Benz noise

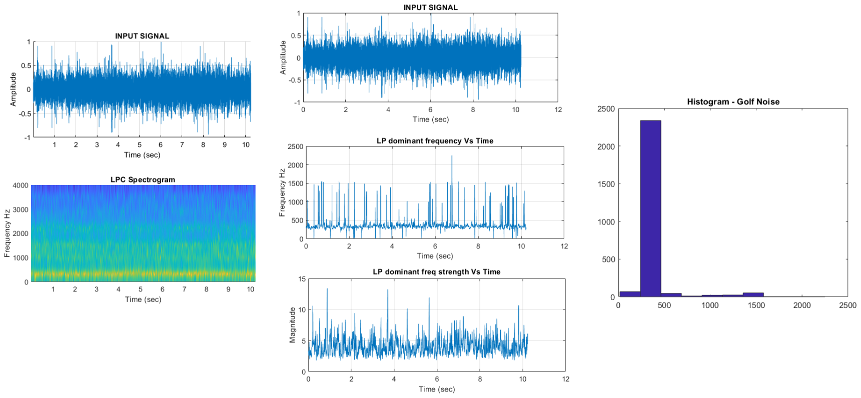

Figure 5. LPC spectrogram, LPC dominant frequency and strength of the dominant frequency for Golf noise

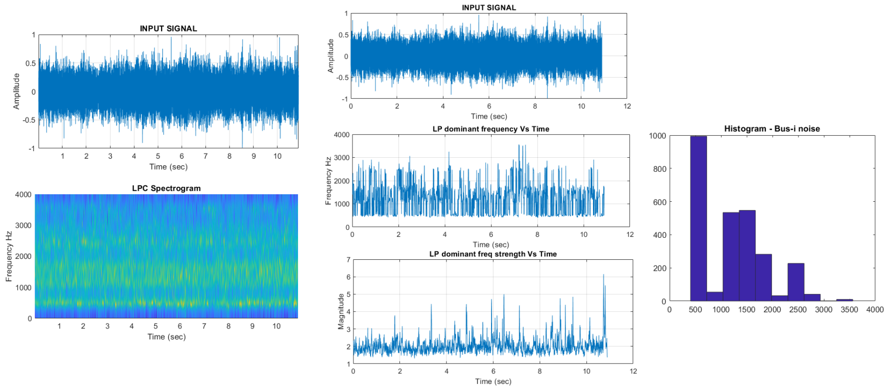

Figure 6. LPC spectrogram, LPC dominant frequency and strength of the dominant frequency for Bus-i noise

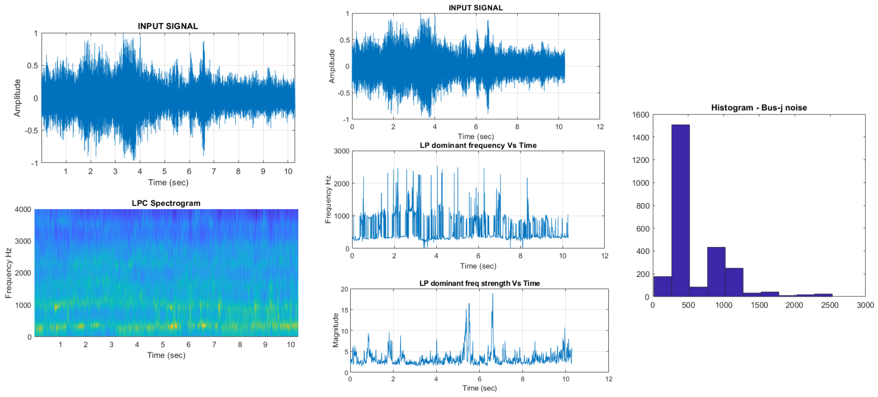

Figure 7. LPC spectrogram, LP dominant frequency and strength of the dominant frequency for Bus-j noise

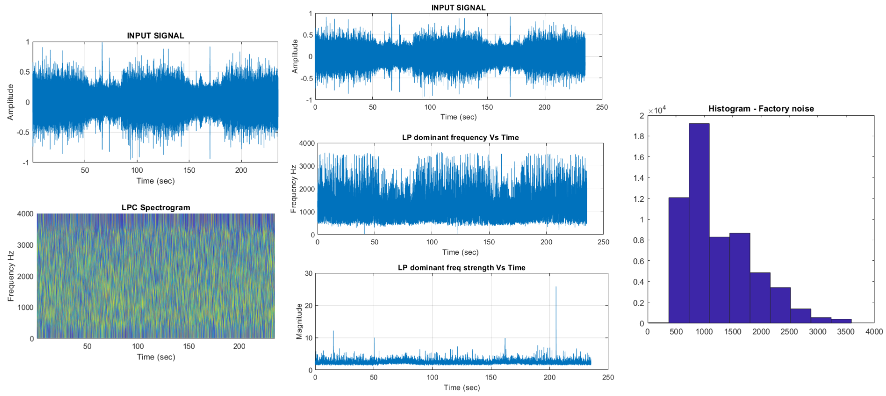

Figure 8. LPC spectrogram, LP dominant frequency and strength of the dominant frequency for Factory noise

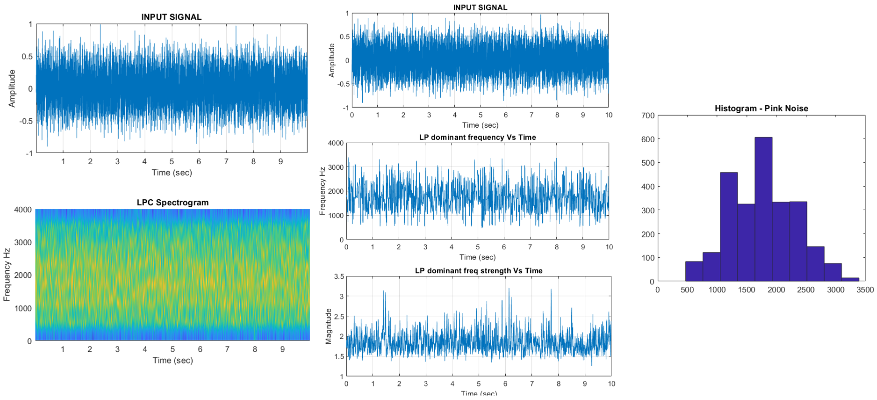

Figure 9. LPC spectrogram, LPC dominant frequency and strength of the dominant frequency for Pink noise

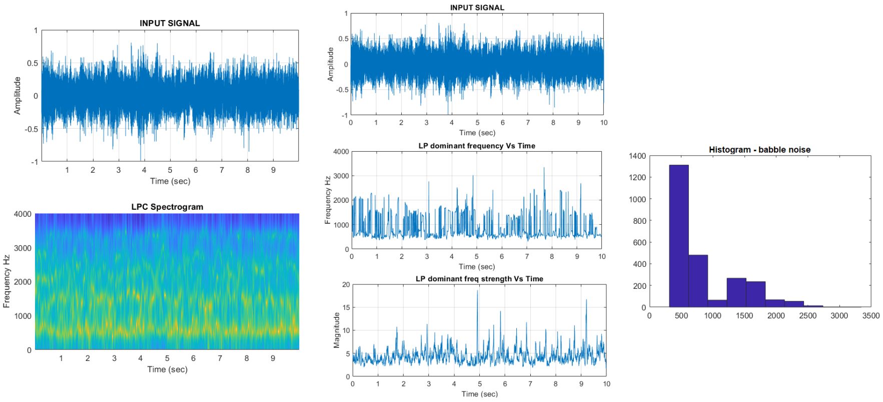

Figure 10. LPC spectrogram, LPC dominant frequency and strength of the dominant frequency for Babble noise

Based on the above analysis, it can be inferred that the low-frequency noises, specifically car, volvo, bus, golf, and benz noises, have a distinct advantage in augmenting dysarthric speech data. Additional analyses like LP spectrogram analysis and Correlation analysis are carried out to support this noise choice.

After the noise signal selection, the next criterion will be the selection of the dB range. Since the augmentation aims to provide more unique samples for speech recognition rather than speech enhancement analysis, only the positive SNR dB ranges are selected in favor of this choice, supporting the new augmented unique examples. In this sense, for the LP analysis, an SNR of +5dB is chosen. The noisy data mentioned above is augmented with the UA dysarthric speech data for the high, moderate, low and very low categories of dysarthric speakers. Figures 11 and 12 show the LP-dominant frequency and dominant frequency strength of noise data augmented onto the low-intelligibility category of dysarthric speakers F02. To study the influence of the noise on the

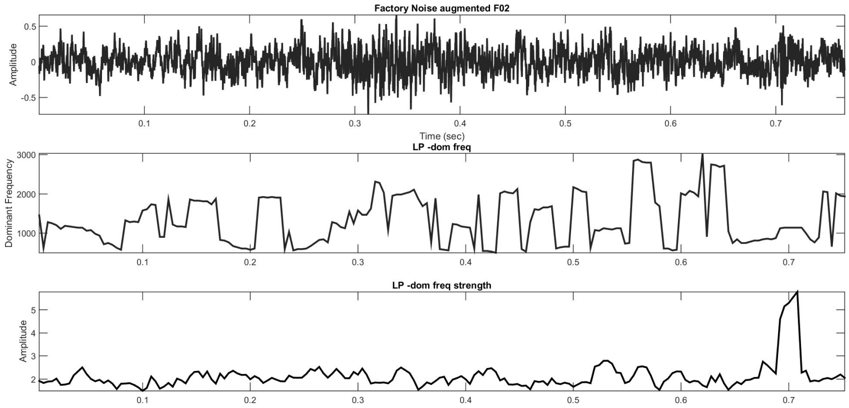

Figure 11. Augmentation of factory noise on F02 - Low-intelligibility category of dysarthric speakers

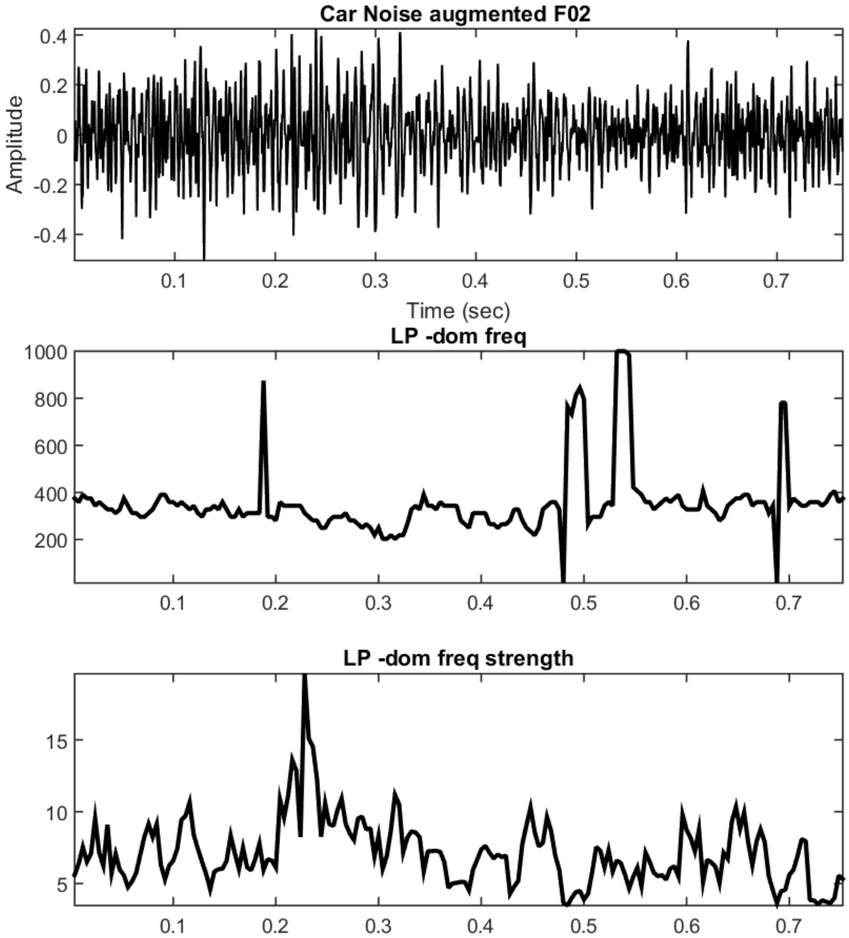

Figure 12. Augmentation of car noise on F02 - Low-intelligibility category of dysarthric speakers

dysarthric speech data, augmentation is performed using the dominant frequency range of the noise.

By analyzing Figures 11 and 12, it is obvious that the impact of car noise on dysarthric speech data is relatively minimal compared to the effect of factory noise. Even with car noise added to the speech, the speech components are still perceivable compared to the factory noise, as shown in Figure 12. The analysis indicates that noise does affect augmentation; however, the crucial factor is the selection of noise from a diverse range of options, as it significantly influences the data augmentation process.

To further validate the noise selection, a correlation analysis is conducted between the clean dysarthric speech data and the noise-augmented dysarthric speech data at +5dB for all categories of noises. Table 1 shows that noises like volvo and car strongly correlate with the dysarthric speech data. Interestingly, even high-frequency noises display positive correlations.

Table 1. Correlation Between Dysarthric Speech Data and Noise Augmented Dysarthric Speech Data

Category of Noise	Correlation coefficient
Pink	0.61
Factory	0.6
Babble	0.81
Car	0.88
Volvo	0.9
Benz	0.87

Therefore, based on this analysis, it can be inferred that low-frequency noises with suitable SNRs are more favorable as a source for data augmentation to generate new training data samples through transformations of the original data. Nevertheless, high-frequency noises can also serve as a potential source for data augmentation, as long as low-frequency segments are carefully selected from those noises to add to the analyzed data.

It is also important to note that when using dysarthric speech data as a source directly to create additional training samples, the speaker's identity and speech errors are preserved in the augmented speech. This ensures that the augmented data retains the identity of the original speaker and maintains the unique characteristics of dysarthric speech without the overpowering noisy data influence.

Section 5 discusses augmenting the dysarthric speech data with the selected noise groups at the appropriate SNR ranges.

5. Augmenting Dysarthric Speech Data with the Selected Noise Groups

From the above section, it is understood that the dominant frequencies of the chosen low-frequency noises have a poor effect when augmenting them with the dysarthric speech data. SNR dB levels ranging from +5dB to +20dB in steps of 5 were taken as SNR level choices while dysarthric speech data is augmented with them.

Based on the previous analysis in Section 4, low-frequency noises like volvo, car, benz, bus-i, bus-j, and golf were chosen for augmentation. Under the five noise conditions at four different SNR levels, the noise data was augmented with the dysarthric speech data at a frame-by-frame level. To ensure that the features of different SNR levels are not just copies of the same original data but examples which are distinct from each other. Figure 13 depicts a line graph for the features extracted using the augmented data at different SNR levels and original speech data to prove the

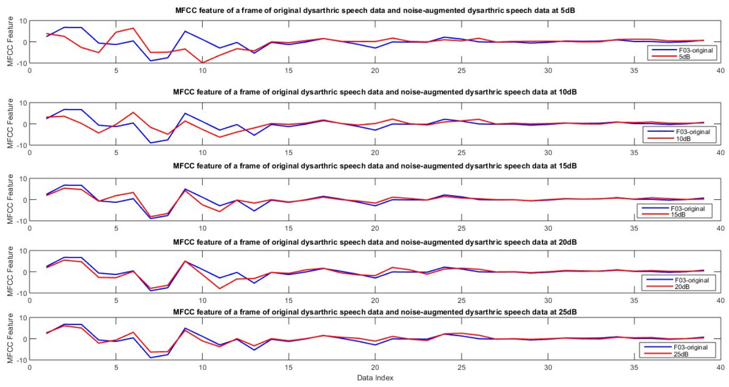

Figure 13. Line Graph comparing the MFCC features of original dysarthric speech data with the features of noise augmented versions

same. The figure shows clearly that the features of the original speech data and the features of the augmented speech data don't lie on each other, indicating they are distinct and hence they can serve as new examples for training. Each dysarthric speech dataset is augmented with the noise data at the varying SNR ranges, giving rise to 24 recent examples (6 noises and 4 db levels) for each original dysarthric speech example. The interesting point of this approach is that unlike other data augmentation approaches, the number of augmented examples is not restricted to a limit, this can be expanded if more ranges of low-frequency noises are included for augmentation.

6. DNN-HMM-based Automatic Dysarthric Speech Recognition System with the Augmented Speech Data

With the augmented dysarthric speech data, a speaker dependent dysarthric speech recognition system is trained for the UA dysarthric speech corpus to verify the performance for noise augmentation both at the level of isolated words and phrases. Using the Noise selection based approach, each example is increased to a number of 24 new ones. The UA corpus has three words per iteration out of which the third word along with its augmented example, is kept for testing. The first two examples and their corresponding augmented example are kept for training. Hence, 50 dysarthric speech datasets (out of which 48 augmented & 2 – original) are kept for training, and 25 dysarthric speech datasets (out of which 24 augmented & 1 original) are kept for testing.

Initially, 39 Mel frequency cepstral coefficient (MFCC) feature vectors (13 static + 13 delta + 13 acceleration) are extracted. Speaker-dependent HMMs are trained using these 50 examples for each word with a varying number of states, varied based on the number of phones in a word, and 2 mixture components per state. With the increased training data, hybrid deep neural network (DNN)-hidden Markov model (HMM)-based isolated word speaker-dependent speech recognition systems are trained for all the 13 dysarthric speakers in UA corpus. HMM is initially trained for the varying resolutions of MFCC. Cepstral mean variance normalization (CMVN) is applied on the extracted features. With the trained HMMs, the DNN is trained using the Kaldi toolkit (Povey et al., 2011). Feature space Maximum Likelihood Linear Regression (fMLLR) transformed features are used as input to the DNN training. The DNN architecture consists of 5 hidden layers using the tanh activation function. DNN training is carried out using 15 epochs for each dysarthric speaker. The performance of each dysarthric speaker is calculated in terms of WER and is tabulated in Table 2.

Table 2. WER performance of UA dysarthric speech recognition systems augmented with the SNR-based noise selection

Speaker ID	Baseline without data augmentation	Car Noise	Benz Noise	Golf Noise	Volvo Noise	Bus-I Noise	Bus-J Noise	WER results Including all 6 noises
M08 (H)	5.97	5.19	5.45	4.21	5.1	5.89	5.12	4.21
M09 (H)	6.49	4.68	5.32	5.32	5.45	4.18	4.21	3.87
M10 (H)	2.34	1.17	0.97	0.95	0.69	1.9	2.01	0.98
M14 (H)	7.2	3.16	2.06	2.48	6.43	5.64	3.42	2.11
F05 (H)	4.42	2.4	2.08	3.46	2.14	3.55	2.32	3.12
F04 (Mod)	16.25	13.21	12.02	16.06	12.02	13.85	12.9	14.21
M05 (Mod)	18.7	12.12	16.82	12.21	16.82	13.46	10.35	14.12
M11 (Mod)	18.79	17.58	15.32	9.7	17.88	8.4	9.35	12.33
F02 (L)	5.45	3.33	4.29	5.12	5.12	4.05	5.41	3.12
M07 (L)	9.09	6.88	8.57	2.95	7.43	4.65	5.08	3.19
F03 (VL)	44.44	40.99	43.12	40.8	31.21	18.23	19.19	32.15
M04 (VL)	98.18	65.18	68.14	43.31	43.12	49.65	43.21	45.33
M12 (VL)	44.85	39.56	40.79	40.61	39.23	30.37	31.64	36.44

Table 2 shows the word error rates (WERs) of the DNN-HMM recognition system trained using noise as a source for data augmentation. Initially, the effect of data augmentation using individual noises is tabulated, and the last column shows the WER of augmenting with all the noises together. It can be observed that compared to each noise taken separately, data augmentation by combining all the noises shows the lowest WER when compared to the baseline system without data augmentation. It can be seen that for high intelligibility dysarthric speakers, noises like golf show a low WER, and for the other categories of dysarthric speakers, noises like car, benz and volvo give a comparatively better reduced WER in relation to the baseline. The noise augmentation in the current work is restricted only to 6 noises. However, it can be extended to any low-frequency noises. Further, the SNR db levels can be extended only up to 20dB, beyond which the sound and the features don’t differ much from the original sample. Hence, additional noises can be included within the SNR levels to increase the example size for dysarthric speech recognition training.

7. Comparison of the Proposed Approach with Those in Recent Data Augmentation Works

The performance of the system proposed by the current work using noise as a source of data augmentation is compared with that of recent dysarthric speech data augmentation approaches in (Thekekara Antony et al., 2023) and (Geng et al., 2022). Both approaches have used UA dysarthric speech corpus as their dysarthric speech data for augmentation.

In (Geng et al., 2022), a comparative investigation is conducted between the approaches like Vocal Tract Length Perturbation (VTLP), tempo and speed adjustments in dysarthric speech data augmentation, and it can be observed that speed perturbation has obtained the lowest WER compared to the baseline system without augmentation. It can be seen from Table 3 that high intelligibility dysarthric speakers have an average WER of 8.04%, which is 6.87% higher than in the proposed work. For the moderate category, the average WER in (Geng et al., 2022) is 17.35%, which is 3.23% higher than in the proposed work. Further, for the low-intelligibility and very low-intelligibility categories of dysarthric speakers, the average WER in (Geng et al., 2022) is 23.93% and 19.45% higher, respectively, than in the proposed work. Also, in comparison with the work of Thekekara Antony et al. (2023), which employed transfer learning approach-based recognition using a virtual microphone and multiresolution feature extraction approach, the approach in the proposed work has obtained the lowest WER, namely of 2.34%, 16.25%, 5.45%, 44.44% for high, moderate, low and very low categories, which is 1.17%, 2.13%, 2.12%, 1.45% higher than in the proposed work for each category, respectively.

In comparison with the previous works, it is interesting to observe that the approach in the current work has obtained the lowest WER

Table 3. Comparison of the results obtained by the approach in the proposed work with those obtained by other two approaches in recent works on data augmentation

Speaker ID	Investigation with speed perturbation approach (Geng et al., 2022)	TL-based ASR system augmented with VM-MRFE (Thekekara Antony et al., 2023)	Proposed Work
M08 (H)	8.04	5.97	5.19
M09 (H)		6.49	5.58
M10 (H)		2.34	1.17
M14 (H)		7.2	3.16
F05 (H)		4.42	2.99
F04 (Mod)	17.35	16.25	15.17
M05 (Mod)		18.7	14.12
M11 (Mod)		18.79	17.47
F02 (L)	27.26	5.45	3.33
M07 (L)		9.09	6.88
F03 (VL)	62.44	44.44	42.99
M04 (VL)		98.18	65.18
M12 (VL)		44.85	43.89

for all the categories of dysarthric speakers, especially the low and very low intelligibility categories. The system in the current work has shown a performance which is very close to that of the approach of Thekekara Antony et al. (2023), clearly depicting that data augmentation approaches using dysarthric speakers' data and Thekekara Antony et al. (2023) shows comparatively better performance in case of data augmentation using normal speaker's speech data. This is because using a dysarthric speaker's speech data for data augmentation retains the identity and originality of the dysarthric speaker. However, in (Thekekara Antony et al. 2023), the number of augmented examples cannot be increased further as in the proposed work due to the poor signal strength while augmenting for an increased array length, as discussed in Section 3.

8. Conclusion

For speakers with dysarthria, speech assistive aids would be a good communication choice. However, developing a speech assistive device using a speech recognition system for dysarthric speakers is challenging due to the sparse availability of dysarthric speech data. To overcome this issue, the technique of data augmentation is used. Data augmentation using dysarthric speech data as a source or normal speakers' data are the two ways of augmenting dysarthric speech data. In the current work, the natural and unintelligible dysarthric speech data is used as a source of data augmentation. Using the dysarthric speech data as a source, helps retain the identity and originality of the dysarthric speakers and their errors. Modifications must be made to turn augmented speech data into natural dysarthric speech data. In the current work, the modification is achieved by adding noise to the natural dysarthric speech data. Noise is an abundant source available for free; when this noise is added to the natural dysarthric speech data, it can be used as various versions of the original data. However, there are limitations regarding the ways it could be augmented with dysarthric speech data. Low-frequency noises are the best sources for augmentation as they would not affect the speech frequency range. Since low-frequency noises are essential for data augmentation, noises like volvo, golf, car, benz, and bus are added to the dysarthric speech data. Initially, 21 examples were in the database, and it has increased to 504 samples (21*6(noises)*4(dB)) of augmented data. Using the augmented data, a DNN-HMM-based dysarthric speech recognition system is trained. It is observed that adding all the noises with the original data and using them for augmentation has reduced the WER greatly for all categories of dysarthric speakers. The obtained WER is also compared with that obtained by recent data augmentation approaches for dysarthric speech data, and the difference was a reduction of up to 44.44% for the data augmentation approach using normal speech and a reduction of up to 2.12% for the data augmentation approach using dysarthric speech data. Further, the number of augmented examples is not restricted to any limit as it can be extended to other low-frequency noises for data augmentation for dysarthric speech.

REFERENCES

Bhat, C., Panda, A. & Strik, H. (2022) Improved ASR Performance for Dysarthric Speech Using Two-stage Data Augmentation. In: *Proceedings of INTERSPEECH 2022, 18-22 September 2022, Incheon, Korea*. pp. 46-50.

Borrie, S. A., Baese-Berk, M., Van Engen, K. & Bent, T. (2017) A relationship between processing speech in noise and dysarthric speech. *The Journal of the Acoustical Society of America*. 141(6), 4660-4667. doi: 10.1121/1.4986746.

Darley, F. L., Aronson, A. & Brown, J. R. (1975) *Motor Speech Disorders*. Philadelphia, Saunders.

Geng, M., Xie, X., Liu, S., Yu, J., Hu, S., Liu, X. & Meng, H. (2022) Investigation of data augmentation techniques for disordered speech recognition. [Preprint] https://arxiv.org/abs/2201.05562.

Jiao, Y., Tu, M., Berisha, V. & Liss, J. M. (2018) Simulating dysarthric speech for training data augmentation in clinical speech applications. In *Proceedings of the 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), 15-20 April 2018, Calgary, Alberta, Canada*. IEEE. pp. 6009-6013.

Jin, Z., Geng, M, Xie, X., Yu, J., Liu, S., Liu, X. & Meng, H. (2021) Adversarial Data Augmentation for Disordered Speech Recognition. In: *Proceedings of INTERSPEECH 2021, 30 August - 3 September 2021, Brno, Czechia. International Speech Communication Association (ISCA)*. pp. 4803-4807.

Kim, H., Hasegawa-Johnson, M., Perlman, A. L., Gunderson, J., Huang, T. S., Watkin, K. & Frame, S. (2008) Dysarthric speech database for universal access research. In: *Proceedings of the 9th Annual Conference of the International Speech Communication Association, INTERSPEECH 2008, 22-26 September 2008, Brisbane, Australia*. pp. 1741-1744.

Menendez-Pidal, X., Polikoff, J. B., Peters, S. M., Leonzio, J. E. & Bunnell, H. T. (1996) The Nemours database of dysarthric speech. In: *Proceedings of the 4th International Conference of Spoken Language Processing, ICSLP '96, 3-6 October 1996, Philadelphia, PA, USA*. IEEE. pp. 1962-1965.

Muthu Philominal, A. J., Nagarajan, T. & Vijayalakshmi, P. (2020) Adaptive multi-band filter structure-based far-end speech enhancement. *IET Signal Processing*. 14(5), 288-299. doi: 10.1049/iet-spr.2019.0226.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G. & Vesely, K. (2011) The Kaldi speech recognition toolkit. In: *Automatic Speech Recognition and Understanding Workshop, 11-15 December 2011, Waikoloa, Hawaii, USA*. IEEE. pp. 1-4.

Rudzicz, F., Namasivayam, A. K. & Wolff, T. (2012) The TORGO database of acoustic and articulatory speech from speakers with dysarthria. *Language Resources and Evaluation*. 46(4), 523-541. doi: 10.1007/s10579-011-9145-0.

Salamon, J. & Bello, J. P. (2017) Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*. 24(3), 279-283. doi: 10.1109/LSP.2017.2657381.

Thekekara Antony, M. C., Nagarajan, T. & Vijayalakshmi, P. (2016) Dysarthric speech corpus in Tamil for rehabilitation research. In: *Proceedings of IEEE Region 10 International Conference (TENCON), 22-25 November 2016, Singapore*. Singapore, IEEE. pp. 2610-2613.

Thekekara Antony, M. C., Nagarajan, T. & Vijayalakshmi, P. (2020) Data Augmentation Using Virtual Microphone Array Synthesis and Multi-Resolution Feature Extraction for Isolated Word Dysarthric Speech Recognition. *IEEE Journal of Selected Topics in Signal Processing*. 14(2), 346-354. doi: 10.1109/JSTSP.2020.2972161.

Thekekara Antony, M. C., Vijayalakshmi, P. & Nagarajan, T. (2023) Data Augmentation Techniques for Transfer Learning-Based Continuous Dysarthric Speech Recognition. *Circuits, Systems, and Signal Processing*. 42(2), 601-622. doi: 10.1007/s00034-022-02156-7.

Vachhani, B., Bhat, C. & Kopparapu, S. K. (2018) Data Augmentation Using Healthy Speech for Dysarthric Speech Recognition. In: *Proceedings of INTERSPEECH 2018, 2-6 September 2018, Hyderabad, India*. International Speech Communication Association (ISCA). pp. 471-475.

Varga, A. & Steeneken, H. J. (1993) 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 Communication*. 12(3), 247-251. doi: 10.1016/0167-6393(93)90095-3.

Xiong, F., Barker, J., Yue, Z. & Christensen, H. (2020) Source domain data selection for improved transfer learning targeting dysarthric speech recognition. In: *Proceedings of the 45th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2020), 4-8 May 2020, Barcelona, Spain*. IEEE. pp. 7424-7428.