An overview of Automatic Speech Recognition Preprocessing Techniques

Conference Paper · March 2022

DOI: 10.1109/DASA54658.2022.9765043

4 authors:

Maria Labied
Université Hassan II de Casablanca
5 PUBLICATIONS **18** CITATIONS
SEE PROFILE

Abdessamad Belangour
71 PUBLICATIONS **567** CITATIONS
SEE PROFILE

Mouad Banane
Université Hassan II de Casablanca
40 PUBLICATIONS **296** CITATIONS
SEE PROFILE

Allae Erraissi
Université Chouaib Doukkali
44 PUBLICATIONS **415** CITATIONS
SEE PROFILE

An overview of Automatic Speech Recognition Preprocessing Techniques

Maria Labied
Laboratory of Information Technology and Modeling LTIM
Hassan II University, Ben M'sik Faculty of Sciences
Casablanca, Morocco
mr.labied@gmail.com

Abdessamad Belangour
Laboratory of Information Technology and Modeling LTIM
Hassan II University, Ben M'sik Faculty of Sciences
Casablanca, Morocco
belangour@gmail.com

Mouad Banane
Laboratory of Artificial Intelligence & Complex Systems Engineering
Hassan II University, Faculty of Legal, Economic and Social Sciences
Casablanca, Morocco
mouad.banane-etu@etu.univh2c.ma

Allae Erraissi
FPSB, Chouaib Doukkali University
El Jadida, Morocco
erraissi.a@ucd.ac.ma

***Abstract*—Speech signal preprocessing is the first and the most important step in the automatic speech recognition process. The preprocessing of speech consists of cleaning the speech signal from ambient and undesirable noises, detecting speech activity, and normalizing the length of the vocal tract. The objective of preprocessing a speech signal is to make the speech recognition systems computationally more efficient through the application of several preprocessing techniques, such as speech pre-emphasis, vocal tract length normalization, voice activity detection, noise removal, framing, and windowing. This paper gives an overview of the fundamentals of speech signal preprocessing techniques, by highlighting the specifics and the requirements of each technique. We also explore all aspects that can improve the results of each technique. We aim that the content of this paper will help researchers improve the quality of their speech recognition systems by identifying appropriate speech preprocessing techniques to use in their experimental settings.**

Keywords— Speech Preprocessing, Automatic Speech Recognition, Pre-Emphasis, Voice Activity Detection, Noise Removal, Speech Enhancement

I. INTRODUCTION

Speech signal processing is an active research field nowadays, the processing of a speech signal consists of multiple phases, at the top of these phases we find speech preprocessing, which is the most important phase in this process and consists of filtering, amplification, noise suppression, speech activity detection, normalization, and speech enhancement. The researches that have been conducted in the field of speech recognition show that the quality of speech recognition depends highly on the efficiency of the speech signal preprocessing[1]. In this paper, we have collected the different speech signal preprocessing techniques used to provide well-prepared data for performing other speech recognition tasks, these processing techniques include Voice Activity Detection, Noise Reduction, Pre-emphasis, Framing, Windowing, and Normalization.

The content of this paper is structured as follows. In Section 2 we review related works concerning the speech preprocessing techniques. In Section 3 we present the different speech preprocessing techniques. In section 4, we illustrate the advantages and disadvantages of the different preprocessing techniques, then we discuss in Section 5 the characteristics of each technique, and finally, we end with our conclusion.

II. RELATED WORK

Speech processing is a core step in automatic speech recognition (ASR), that has been extensively researched; Therefore, we present below some related works addressing speech signal preprocessing techniques in specific contexts. Multiple works were dedicated to investigating speech denoising techniques, a deep learning-based approach has been proposed by Alamdari et al. [2] to improve denoising real-world speech, without requiring pre-cleaned speech signals. The experimental results of this research have proved that the proposed deep-learning denoising approach outperforms the conventional supervised denoising approach. Excluding non-speech segments before denoising a speech signal represented a good alternative to minimize computational costs of deep learning denoising tasks, on this basis, lee et al. [3] have presented a speech denoising strategy based on speech segment detection before starting the denoising process. The obtained results showed that the accuracy of the proposed denoising strategy is similar to or even better than that of the Wavenet-based denoising method[4], which is one of the recent deep neural networks based denoising methods. The Advances in deep learning-based methods have improved the performance of speech denoising methods, a Denoising Autoencoder with a Multi-Branched Encoder (DAEME)[5] has been proposed to deal with the local focus on specific noisy conditions and mismatched noisy conditions during testing.

Next to speech denoising, Voice Activity Detection (VAD) is considered a preliminary phase in speech processing. The performance of speech processing, speech recognition, speech enhancement., highly depends on the performance of VAD output. A robust background noise VAD algorithm has been introduced by Xu et al[6] to distinguish between voiced and unvoiced parts of speech, and to determine the absence or presence of speech. Also, Singh et al [7] have implemented a Digital signal processors-based VAD system and a noises suppression algorithm. The experiments of this research show that the signal-to-noise ratio (SNR) value was significantly increased when noise suppression algorithm and VAD were used. The VAD was considered a crucial task of many speech recognition systems, it was the heart of a Voice Operated Switch(VOS) aircraft application[8]. In this works a reliable, VAD scheme for VOS application was explored by comparing artificial neural

network-based detectors, fuzzy logic detectors, and linear Energy-based detectors. Also, VAD has been applied as a core activity to provide voice-based home appliances. Jat et al [9] have presented an easy-to-integrate VAD-based home automation system, to improve voice commands automation processes.

The choice of appropriate windowing techniques, therefore, plays an essential role in signal processing. Many studies have been carried out to find the best window function to use in the process of windowing the speech signal. A comparative study of different window functions was presented by Aparna et al [10], which showed that triangular and hamming windows perform better than other windowing functions. Also, a new method has been proposed by Gaffar et al [11] to improve the stability of the signal characteristics to minimize the significant changes that occur in the signal during the windowing step.

III. Speech Signal Preprocessing Techniques

The preprocessing of the speech signal is mainly performed to make the speech signal ready for the feature extraction speech recognition phase. The speech signal is preprocessed in these ordered steps. In the first step, the VAD is executed to keep only the voice parts of a speech signal. Then, noise is reduced to allow balancing the frequency of signal spectrum by applying the pre-emphasis of the speech signal, after that, the framing of the whole signal is applied. Then, a windowing function and a normalization of the vocal tract are performed to improve the signal spectrum and SNR of the speech signal. The specifics of each preprocessing step are detailed in the sections below.

A. Voice activity detection (VAD)

VAD is an important technique in speech signal preprocessing[8], in which the voiced parts of speech are identified. In ASR speech activity detection is used to determine the start and the end of speech utterances, also it allows to meet the limited hardware capabilities and to optimize the CPU consumption.

The application of VAD consists of dividing the speech signal into short frames, then determining whether the frame contains speech or not. The VAD algorithms are based on the selection of features that represent the discriminative properties of speech and noise. Two categories of features could be distinguished, the first category is the time-domain features such as Short-Term Energy (STE), Zero-Crossing Rate (ZCR), and Short-Time Average Magnitude (STAM), are between the widely used features, due to their simplicity. However, these features are degradable by background noise. The second category is the frequency-domain features, such as the Spectral Flatness (SF) and Spectral Power (SP). Other features simulate human perception to detect the presence of speech, such as the Amplitude Modulation Spectrogram (AMS) and Spectro-Temporal Modulation (STM).

B. Noise removal

Environment noise plays an important role in speech preprocessing, denoising a speech signal improves the quality of speech and enhances the robustness of speech recognition systems. Noise removal is the process of cleaning the noise from a mixed sound of speech and noise to keep only the clean speech[3].

C. Pre-emphasis

The frequency components of a speech signal fall in the high frequencies, the pre-emphasis is applied to flatten the magnitude spectrum and balance the high and low-frequency components of the speech signal. On the other hand, pre-emphasis filtering is used for reducing the high dynamic range of speech waveform and enhancing the signal-to-noise ratio. The pre-emphasis preprocessing method is less used in speech recognition processing. However, this preprocessing technique increases the energy of the speech signal at high frequencies, which affects the consistency of the resulting spectrum among frames[12].

D. Framing

Since the speech signal has a non-stationary nature, the properties of a signal are not static over time, However, in short time intervals, the speech signal is considered as a stationary signal and easy to process[13]. The Framing step fills this gap by splitting the continuous speech signal into a series of blocks called frames of 20-40ms length to allow block processing of the speech signal[14][11]. Framing is a fundamental technique in speech signal processing, the obtained frames have an equal length and are stationary over time, which makes extracting useful properties of the speech signal much easier.

E. Windowing

Windowing is the analysis process of speech signals in each frame, which consists of multiplying a speech signal waveform by a time window function, to emphasize predefined properties of the speech signal[14][11][15].

The concept of windowing is primarily used to help smooth the signal and avoid signal discontinuity produced by the spectral distortion during the framing stage, by reducing the signal at the beginning and end of each frame to zero. However, the use of windowing techniques sometimes has an impact on changing the signal at the beginning and end of the frame. Multiplying the speech wave by the window function has two main impacts. Firstly, to avoid a sharp change at the endpoints of a frame it gradually reduces the amplitude at both ends of the time interval. The second effect is to convolve the speech spectrum and the Fourier transform of the window function[15]. Thus, the choice of the window function to be applied to reduce spectral distortion must satisfy two conditions: high-frequency resolution and low spectral leakage. In ASR analysis it is recommended to use a short window where the frame length is between 20 and 30ms, with overlapping between 5 and 10ms [14].

F. Normalization

The normalization of a speech signal is the final step in speech signal preprocessing, which involves balancing the signal spectrum and transforming the signal data into a normalized form based on a threshold [16]. Normalization is mainly used of reducing the noise impact, speech signal distortion, and channel distortion [17]. In the normalization phase, SNR is improved and the signal spectral variation is normalized or eliminated. There exist multiple normalization techniques [18][17], we explore the specifics of each technique in sections below.

IV. Preprocessing techniques Advantages and Disadvantages

A. Noise removal methods

Three categories of noise removal methods can be distinguished:

1) Statistical feature-based methods

Are the earliest used methods for speech denoising, an example of these methods are the Wiener filtering[19], the Spectral Subtraction[20][21][22], the wavelet transformation denoising[23], and non-negative matrix factorization[24].

2) Denoising Auto-Encoder(DAE)-based methods

DAE was used for the first time in image processing to denoise features for classification[25]. In recent years, the auto-Encoder denoising concept has been adopted for extracting clean speech from noisy speech data. DAE is used in various ways for denoising speech, like speaker-aware DAE[26], time-domain convolutional DAE[27], and multi-branched encoder DAE[5].

3) Deep neural network-based models

The concept of these methods consists in applying DNNs to model the non-linear relationship between clean and noisy speech signals [28]. Various DNN denoising models have been proposed, In [29] a Convolutional Neural Network (CNN) denoising model was proposed, also an end-to-end Wavenet-based denoising model was proposed by google show better performance for denoising speech, this models outperform the most widely used Wiener filter method [4]. Even though the requirement and the higher computing costs, deep learning-based methods show high performance. Several pieces of research have investigated the efficient deep learning denoising models while reducing the computational costs, and at the same time enhance the real-time speech denoising[30][2].

B. Windowing Functions

Windowing involves the application of a window function which states that if a certain interval is chosen, it returns a non-zero finite value in the interval and a zero value outside the interval. There exist several window functions[31], for spectral performance analysis and digital finite impulse response filter, which can be divided into two groups:

- *Fixed window functions:* rectangular window, Hann window, Hamming window, Blackman window...
- *Adjustable window functions*[32]*:* Kaiser window, Saramäki window, ultraspherical window...

Table.1 summarizes the different characteristics and the constraints of the most commonly used windowing functions

TABLE I. CHARACTERISTICS AND CONSTRAINTS OF WINDOWING FUNCTIONS

Window function	Characteristics	Constraints
Rectangular	-Perfectly spaced Periodic signals. -Signals length shorter than the window length. -Provide the exact frequency of the peaks of a signal	- Results in leakage in present discontinuities. - lower side lobe attenuation
Triangular	-Exhibits a non-negative Fourier transform - self-convolutional Simplicity when computing coefficient Lower attenuation -Perform better at a low SNRs	-low energy concentration in the sides lobe
Hann	- Higher stopband attenuation - Sharper fall. - Fast increases the stopband attenuation of the following lobes.	- Wider main lobe. - Distorting. The waveform spectrum
Hamming	- Transition Region Between 3-41db. - Better Selectivity For Large Signals. - Wider Transition Region. - Higher Stopband Attenuation.	- Rest of the side lobes are higher Discontinuous at the edges.
Blackman	- High sides lobe attenuation [10]. - Greater stopband attenuation is easy to obtain. -smoothly taper a signal at its edges.	-Wider main lobe.
Flat-top	- Performing calibration. - Exact amplitude of the signal. - Exhibiting the best resolution of amplitude. - Better amplitude accuracy in the frequency domain. - Employed on data where frequency peaks are distinct.	- A greater amount of frequency domain leakage.

C. Normalization functions

1) Cepstral Mean Normalization (CMN)

CMN is the widely used technique for compensating the variability of the speech signal in the cepstral domain[18]. This technique is considered the simplest to implement and the most used normalization technique for large vocabulary ASR applications. The main role of CMN is to remove the convolutional distortions caused by the environmental characteristics of speech(recording device, noises …) [33], at the same time it reduces the effect of speech style. However, the CMN fails to remove convolutional distortions when their time constant approaches the length of the analysis window[34].

2) Cepstral Variance Normalization (CVN)

CVN is a similar technique to CMN, which consists of estimating the variance of each cepstral dimension and then normalizing it into a unity[35]. CVN is not associated with any particular type of distortion. However, it does offer some robustness against speaker variability, additive noise, and acoustic channels[34].

3) Cepstral Mean and Variance Normalization (CMVN)

CVN and CMN are often paired together as the cepstral mean and variance normalization. Therefore, CMVN benefits from both the sample mean and standard deviation, to normalize the cepstral sequence. The use of CMVN positively affects the accuracy and reduces the error rate when using a clean or multi-style acoustic model. Unlike CMN, CMVN shifts and scales the energy term [34].

4) Histogram Equalization (HEQ)

HEQ transforms the cepstral coefficients by normalizing the probability density function and making it equal to the reference probability density function [36]. HEQ is frequently used in Digital Image Processing but has become an important normalization technique for robust speech processing[37]. Compared to CMNV, HEQ doesn't only eliminate the linear effects of noises but also eliminates the non-linear distortions caused by noise[38].

5) Cepstral Gain Normalization (CGN)

CGN is similar to CVN and MVN. However, CVN and MVN are not sufficient to identify the effects of additive noise[33]. The philosophy of CGN is based on the approximate model, which makes this normalization method robust to noise. The application of CGN starts by Subtracting the average of cepstral coefficients which is known as CMN. Then Normalize gain to unity by calculating the maximum and the minimum sample values of cepstral dimension coefficients[35]. The main benefit of using CGN is the gain of noise-robust performance whatever the environment's noise and without distorting the original speech signal.

6) Quantile-based Cepstral dynamics Normalization (QCN)

QCN is a cepstral dynamics normalization technique inspired by CVN, CMN, and CGN techniques, which is mainly used to reduce the sensitivity of the normalization to outliers [39][33]. QCN uses the quantile estimates for each cepstral dynamic dimension. Compared to the other normalization techniques QCN gives an accurate alignment of the cepstral distribution of samples and proves robustness to distributions shapes changes and reverberations[40].

7) Wiener-filtering (WF)

WF is a normalization technique that is used to remove the additive noise on the basics of wiener gains[41][3]. WF considers that the speech and noises are not correlated and minimizes the mean square error. A side of the advantages of WF, there is some weakness to be highlighted, WF estimate the power of clean signal and noise before filtering ignoring the non-stationary nature of speech, also it considers gain function as fixed at every frequency[21].

Table 2 provides a full overview of the advantages and disadvantages of the pre-listed normalization techniques CMN, CVN, CMVN, HEQ, CGN, QCN, and WF.

TABLE II. ADVANTAGES AND DISADVANTAGES OF NORMALIZATION TECHNIQUES

Normalization Technique	Advantages	Disadvantages
CMN	- Simplest implementation [34] - Reduce the effect of speech style [33]	- Less effective to remove the convolutional distortions
CVN	- Robust against additive noises - independent from distortion types [34]	- Not sufficient to identify the effects of additive noises
CMVN	- Eliminates linear effects of noise	- Unable to eliminate the non-linear distortions caused by noise
HEQ	- Computational inexpensive [36]. - Low storage costs [36]. - Eliminate the non-linear distortions noise type and SNR independent	- Require a large amount of data - Dependency to Cumulative density function calculation
CGN	-High noise-robust performance	-
QCN	- Accurate alignment of the cepstral distributions. - Robustness to distributions shapes change Robustness to additive noise and reverberation.	-
WF	- Removing additive noise from a speech signal. - Optimal mean square error.	- Ignoring the non-stationary nature of speech signals. - Estimating signal power before filtering.

V. DISCUSSION

The different preprocessing techniques used in ASR are discussed above. This leads to an understanding of a few very important methods and techniques to be applied for successful speech recognition. First, to ensure a good entry point to the preprocessing workflow of a speech signal, VAD represents the key step in this workflow, as we have mentioned the frequency-domain features-based VAD methods are the best choice as they can simulate human perception to detect the presence of speech. Second, clear noises from the speech signal offer core part of the preprocessing workflow, the noise removal converge from Statistical feature-based methods and Denoising Auto-Encoder (DAE)-based methods to Deep neural network-based models for speech signal denoising. These recent deep learning methods have significantly improved the quality of speech denoising[42][41][3]. Also, In recent years VAD start to be merged within the noise removal step as was mentioned by Muzammel et al[22]. Also, the VAD has benefited from the deep learning advances and multiple approaches have been proposed [43][8], to make an important improvement in VAD performance.

After detecting speech segments, clearing them from noises, emphasizing higher frequencies, and splitting the whole speech signal into frames it still prepares the speech signal for features extraction. In the windowing step, various windowing functions are incorporated in the processing of the speech signal, the rectangular, Hamming, Hann, and Blackman windows are the most popular and commonly used windowing functions. The Rectangular window function can perfectly space periodic signals and provide the exact frequency of the peaks of a signal, but it results in leakage in present discontinuities. The Hamming window function has a better selectivity For Large Signals and a Wider Transition Region but it results in a higher discontinuity at the edges. With the Hann window function, we can obtain a higher stopband attenuation, however, this window function may lead to distorting the waveform spectrum. We can deduce that no window function is the best in all aspects, rather the selection of the windowing function should be based on speech recognition requirements.

The Normalization of speech signal depends highly on the output of the applied windowing function. The pre-listed normalization techniques CMN, CVN, HEQ, CGN, QCN, and WF have been reviewed in this paper in terms of their behaviors, the advantages, and the disadvantages of each. The CMN is the simplest normalization technique to implement and it is effective for reducing the speech style effects, but it is less performant to remove convolutional distortions. The CVN is known for its robustness against additive noises, but it is not sufficient to identify their effects. The HEQ is one of the optimal normalization techniques, it permits low storage costs, eliminates the non-linear distortions, and is computationally inexpensive, however, it requires a large amount of data. The CGN is one of the normalization techniques that ensure high noise-robust performance. The QCN gives an accurate alignment of the cepstral distributions. while WF results in an optimal mean square error but ignores the non-stationary nature of speech signals and estimate signal

power before filtering. From the listed advantages and disadvantages of each windowing technique, we can see that the application of any normalization technique should consider the noisy conditions of the speech signal and the non-stationary nature of speech signals

VI. CONCLUSION

In this work, we presented an overview of commonly used speech signal preprocessing techniques used in ASR. Since speech signal preprocessing affect highly the quality of ASR, choosing the appropriate techniques in each of the preprocessing steps should be considered the most crucial decision. In this review, we tried to summarize some of the strengths and weaknesses of each preprocessing technique discussed above. The finding of this paper has shown that the effectiveness and the quality of speech recognition highly depend on the output of the preprocessing of the speech signal steps, in particular in noise removal, VAD, Windowing, and normalization. In our future work, we will investigate the preprocessing of Moroccan dialect "DARIJA" speech to prepare the speech dataset for the next steps of the speech recognition process.

REFERENCES

[1] N. Mamatov, N. Niyozmatova, and A. Samijonov, "Software for preprocessing voice signals," *Int. J. Appl. Sci. Eng.*, vol. 18, no. 1, pp. 1–8, Mar. 2021, doi: 10.6703/IJASE.202103_18(1).006.

[2] N. Alamdari, A. Azarang, and N. Kehtarnavaz, "Improving deep speech denoising by Noisy2Noisy signal mapping," *Appl. Acoust.*, vol. 172, 2021, doi: 10.1016/j.apacoust.2020.107631.

[3] S. Lee and H. Kwon, "A Preprocessing Strategy for Denoising of Speech Data Based on Speech Segment Detection," pp. 1–24, 2020, doi: 10.3390/app10207385.

[4] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2018-April, pp. 5069–5073, 2018, doi: 10.1109/ICASSP.2018.8462417.

[5] C. Yu *et al.*, "Speech Enhancement Based on Denoising Autoencoder with Multi-Branched Encoders," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 2756–2769, 2020, doi: 10.1109/TASLP.2020.3025638.

[6] N. Xu, C. Wang, and J. Bao, "Voice activity detection using entropy-based method," in *2015 9th International Conference on Signal Processing and Communication Systems (ICSPCS)*, Dec. 2015, pp. 1–4, doi: 10.1109/ICSPCS.2015.7391751.

[7] C. Singh, M. Venter, R. K. Muthu, and D. Brown, "A Real-Time DSP-Based System for Voice Activity Detection and Background Noise Reduction," in *Intelligent Speech Signal Processing*, Elsevier, 2019, pp. 39–54.

[8] Bharath Y.K, Veena S, Nagalakshmi K.V, M. Darshan, and R. Nagapadma, "Development of robust VAD schemes for Voice Operated Switch application in aircrafts: Comparison of real-time VAD schemes which are based on Linear Energy-based Detector, Fuzzy Logic and Artificial Neural Networks," in *2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*, 2016, no. 1, pp. 191–195, doi: 10.1109/ICATCCT.2016.7911990.

[9] D. S. Jat, A. S. Limbo, and C. Singh, "Voice Activity Detection-Based Home Automation System for People With Special Needs," in *Intelligent Speech Signal Processing*, Elsevier, 2019, pp. 101–111.

[10] R. Aparna and P. L. Chithra, "Role of Windowing Techniques in Speech Signal Processing For Enhanced Signal Cryptography," in *Advanced Engineering Research and Applications*, 2017, pp. 446–458.

[11] A. F. Onnilita Gaffar, R. Malani, Supriadi, A. Wajiansyah, and A. B. Wicaksono Putra, "A multi-frame blocking for signal segmentation in voice command recognition," in *2020 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, Jul. 2020, pp. 299–304, doi: 10.1109/ISITIA49792.2020.9163761.

[12] R. Vergin and D. O'Shaughnessy, "Pre-emphasis and speech recognition," *Can. Conf. Electr. Comput. Eng.*, vol. 2, pp. 1062–1065, 1995, doi: 10.1109/ccece.1995.526613.

[13] O. K. Hamid, "Frame Blocking and Windowing Speech Signal," *J. Information, Commun. Intell. Syst.*, vol. 4, no. 5, pp. 87–94, 2018.

[14] Y. A. Ibrahim, J. C. Odiketa, and T. S. Ibiyemi, "Preprocessing technique in automatic speech recognition for human computer interaction: an overview," *Ann. Comput. Sci. Ser.*, vol. 15, no. 1, pp. 186–191, 2017.

[15] S. Furui, *Digital Speech Processing, Synthesis and Recognition*, vol. 148. 2014.

[16] M. M. Hasan, H. Ali, M. F. Hossain, and S. Abujar, "Preprocessing of Continuous Bengali Speech for Feature Extraction," *2020 11th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2020*, pp. 1–4, 2020, doi: 10.1109/ICCCNT49239.2020.9225469.

[17] R. Singh, U. Bhattacharjee, and A. K. Singh, "Performance Evaluation of Normalization Techniques in Adverse Conditions," *Procedia Comput. Sci.*, vol. 171, no. 2019, pp. 1581–1590, 2020, doi: 10.1016/j.procs.2020.04.169.

[18] O. Kalinli, G. Bhattacharya, and C. Weng, "Parametric Cepstral Mean Normalization for Robust Speech Recognition," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2019-May, pp. 6735–6739, 2019, doi: 10.1109/ICASSP.2019.8683674.

[19] S. Dr.China Venkateswarlu, Ks. Prasad, As. Reddy, Sc. Venkateswarlu α, Ks. Prasad Ω, and As. Reddy β, "Improve Speech Enhancement Using Weiner Filtering," *Glob. J. Comput. Sci. Technol.*, vol. 11, no. 7, 2011.

[20] S. V. Vaseghi, "Spectral Subtraction," in *Advanced Signal Processing and Digital Noise Reduction*, Wiesbaden: Vieweg+Teubner Verlag, 1996, pp. 242–260.

[21] N. Upadhyay and A. Karmakar, "Speech Enhancement using Spectral Subtraction-type Algorithms: A Comparison and Simulation Study," *Procedia Comput. Sci.*, vol. 54, pp. 574–584, 2015, doi: 10.1016/j.procs.2015.06.066.

[22] C. S. Muzammel, M. Hasan, K. Ahammad, and M. H. Mukti, "Noise Reduction from Speech Signals using Modified Spectral Subtraction Technique," *Eur. J. Eng. Res. Sci.*, vol. 3, no. 7, p. 78, 2018, doi: 10.24018/ejers.2018.3.7.838.

[23] C. P. Dautov and M. S. Ozerdem, "Wavelet transform and signal denoising using Wavelet method," *26th IEEE Signal Process. Commun. Appl. Conf. SIU 2018*, pp. 1–4, 2018, doi: 10.1109/SIU.2018.8404418.

[24] S. Vanambathina, "Speech Enhancement Using an Iterative Posterior NMF," *New Front. Brain - Comput. Interfaces*, pp. 1–18, 2020, doi: 10.5772/intechopen.84976.

[25] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, and L. Bottou, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion.," *J. Mach. Learn. Res.*, vol. 11, no. 12, 2010.

[26] F.-K. Chuang, S.-S. Wang, J. Hung, Y. Tsao, and S.-H. Fang, "Speaker-Aware Deep Denoising Autoencoder with Embedded Speaker Identity for Speech Enhancement," in *Interspeech 2019*, Sep. 2019, pp. 3173–3177, doi: 10.21437/Interspeech.2019-2108.

[27] N. Tawara, T. Kobayashi, and T. Ogawa, "Multi-Channel Speech Enhancement Using Time-Domain Convolutional Denoising Autoencoder," in *Interspeech 2019*, Sep. 2019, pp. 86–90, doi: 10.21437/Interspeech.2019-3197.

[28] A. Kumar and D. Florencio, "Speech enhancement in multiple-noise conditions using deep neural networks," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 08-12-Sept, pp. 3738–3742, 2016, doi: 10.21437/Interspeech.2016-88.

[29] H. Zhao, S. Zarar, I. Tashev, and C. H. Lee, "Convolutional-Recurrent Neural Networks for Speech Enhancement," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2018-April, pp. 2401–2405, 2018, doi: 10.1109/ICASSP.2018.8462155.

[30] S. Sonning, C. Schuldt, H. Erdogan, and S. Wisdom, "Performance Study of a Convolutional Time-Domain Audio Separation Network for Real-Time Speech Denoising," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2020-May, pp. 831–835, 2020, doi: 10.1109/ICASSP40776.2020.9053846.

[31] K. M. Prabhu, *Window Functions and Their Applications in Signal Processing*. CRC Press, 2014.

[32] A. Datar, A. Jain, and P. C. Sharma, "Design and performance analysis of adjustable window functions based cosine modulated filter banks," *Digit. Signal Process.*, vol. 23, no. 1, pp. 412–417, Jan. 2013, doi: 10.1016/j.dsp.2012.07.007.

[33] D. Grozdić, S. Jovičić, D. Š. Pavlović, J. Galić, and B. Marković,

"Comparison of cepstral normalization techniques in whispered speech recognition," *Adv. Electr. Comput. Eng.*, vol. 17, no. 1, pp. 21–26, 2017, doi: 10.4316/AECE.2017.01004.

[34] J. Droppo and A. Acero, "Environmental Robustness," in *Springer Handbook of Speech Processing*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 653–680.

[35] S. Yoshizawa, N. Hayasaka, N. Wada, and Y. Miyanaga, "Cepstral gain normalization for noise robust speech recognition," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. I-209–12, doi: 10.1109/ICASSP.2004.1325959.

[36] A. de la Torre, A. M. Peinado, J. C. Segura, J. L. Perez-Cordoba, M. C. Benitez, and A. J. Rubio, "Histogram equalization of speech representation for robust speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 3, pp. 355–366, May 2005, doi: 10.1109/TSA.2005.845805.

[37] Z. Huimin, J. Xupeng, and L. Dongmei, "An Iterative Post-processing Approach for Speech Enhancement," in *Proceedings of the 2019 4th International Conference on Multimedia Systems and Signal Processing - ICMSSP 2019*, 2019, pp. 130–134, doi: 10.1145/3330393.3330427.

[38] Y.-C. Kao and B. Chen, "Distribution-based feature normalization for robust speech recognition leveraging context and dynamics cues," in *Interspeech 2013*, Aug. 2013, pp. 2958–2962, doi: 10.21437/Interspeech.2013-269.

[39] H. Boril and J. Hansen, "Unsupervised equalization of Lombard effect for speech recognition in noisy adverse environment," 2009, pp. 3937–3940, doi: 10.1109/ICASSP.2009.4960489.

[40] A. Boulmaiz, D. Messadeg, N. Doghmane, and A. Taleb-Ahmed, "Robust acoustic bird recognition for habitat monitoring with wireless sensor networks," *Int. J. Speech Technol.*, vol. 19, no. 3, pp. 631–645, Sep. 2016, doi: 10.1007/s10772-016-9354-4.

[41] L. Wang, W. Zheng, X. Ma, and S. Lin, "Denoising Speech Based on Deep Learning and Wavelet Decomposition," *Sci. Program.*, vol. 2021, pp. 1–10, Jul. 2021, doi: 10.1155/2021/8677043.

[42] A. Azarang and N. Kehtarnavaz, "A review of multi-objective deep learning speech denoising methods," *Speech Commun.*, vol. 122, no. February, pp. 1–10, 2020, doi: 10.1016/j.specom.2020.04.002.

[43] C. Yu, K.-H. Hung, I.-F. Lin, S.-W. Fu, Y. Tsao, and J. Hung, "Waveform-based Voice Activity Detection Exploiting Fully Convolutional networks with Multi-Branched Encoders," no. March 2021, 2020, [Online]. Available: http://arxiv.org/abs/2006.11139.