Continuous Wavelet Transform for Severity-Level Classification of Dysarthria

Aastha Kachhi(✉), Anand Therattil, Priyanka Gupta, and Hemant A. Patil

Speech Research Lab, DA -IICT, Gandhinagar 382007, Gujarat, India
{aastha_kachhi,anand_therattil,priyanka_gupta,hemant_patil}@daiict.ac.in

Abstract. Dysarthria is a neuro-motor speech defect that causes speech to be unintelligible and is largely unnoticeable to humans at various severity-levels. Dysarthric speech classification is used as a diagnostic method to assess the progression of a patient's severity of the condition, as well as to aid with automatic dysarthric speech recognition systems (an important assistive speech technology). This study investigates the significance of Generalized Morse Wavelet (GMW)-based scalogram features for capturing the discriminative acoustic cues of dysarthric severity-level classification for low-frequency regions, using Convolutional Neural Network (CNN). The performance of scalogram-based features is compared with Short-Time Fourier Transform (STFT)-based features, and Mel spectrogram-based features. Compared to the STFT-based baseline features with a classification accuracy of 91.76%, the proposed Continuous Wavelet Transform (CWT)-based scalogram features achieve significantly improved classification accuracy of 95.17% on standard and statistically meaningful UA-Speech corpus. The remarkably improved results signify that for better dysarthric severity-level classification, the information in the low-frequency regions is more discriminative, as the proposed CWT-based time-frequency representation (scalogram) has a high-frequency resolution in the lower frequencies. On the other hand, STFT-based representations have *constant* resolution across all the frequency bands and therefore, are not as better suited for dysarthric severity-level classification, as the proposed Morse wavelet-based CWT features. In addition, we also perform experiments on the Mel spectrogram to demonstrate that even though the Mel spectrogram also has a high frequency resolution in the lower frequencies with a classification accuracy of 92.65%, the proposed system is better suited. We see an increase of 3.41% and 2.52% in classification accuracy of the proposed system to STFT and Mel spectrogram respectively. To that effect, the performance of the STFT, Mel spectrogram, and scalogram are analyzed using $F1$-Score, Matthew's Correlation Coefficients (MCC), Jaccard Index, Hamming Loss, and Linear Discriminant Analysis (LDA) scatter plots.

Keywords: Wavelet transform · Dysarthria · UA-Speech corpus · Morse wavelet · CNN

S. R. M. Prasanna et al. (Eds.): SPECOM 2022, LNAI 13721, pp. 312–324, 2022.
https://doi.org/10.1007/978-3-031-20980-2_27

1 Introduction

Proper coordination between brain and speech-producing muscles is required for the production of speech sounds [15]. Lack of this coordination leads to speech disorders, such as aparaxia, dysarthria, and stuttering. These disorders affect a person's ability to produce speech sounds. They are further categorized as neurological or neurodegenerative diseases, such as cerebral palsy or Parkinson's disease. The severity-level of these diseases might be mild or severe, depending upon the impact on the area of the brain. In the case of mild severity, the patient may mispronounce a few words, whereas, in high severity, the patient lacks the ability to produce intelligible speech. Among these speech disorders, dysarthria is a relatively common speech disorder [24]. Dysarthria is a neuromotor speech disorder. The muscles that produce speech are weak in people with this disorder. Dynamic movements of articulators, such as lips, tongue, throat, and upper respiratory tract system are also affected due to brain damage. Apart from brain damage, cerebral palsy, muscular dystrophy, and stroke are also some of the other factors, which can cause dysarthria [19].

Severity-level of dysarthria depends on the impact and damage to the area of neurological injury, which is diagnosed using a brain and nerve test. The type, underlying cause, severity-level, and its symptoms, all influence the manner in which it is treated [4]. Due to this uncertainty in treatment, researchers are motivated to develop speech assistive tools for dysarthric intelligibility categorization.

In the literature, dysarthria severity-level classification has been exploited extensively using Short-Time Fourier Transform (STFT) [9], and various acoustical features [1]. State-of-the-art feature sets, such as Mel Frequency Cepstral Coefficients (MFCC) feature set was employed in [12] due to its capacity of capturing global spectral envelope properties. In addition to a perceptually-motivated state-of-the-art feature set, glottal excitation source parameters derived from the quasi-periodic sampling of the vocal tract system were implemented in [8]. In the signal processing framework, due to the wide and dynamic range of multiple frequency components in short-time spectra, speech signals are considered to be non-stationary signals. Due to the dynamic movements of articulators, the frequency spectrum varies instantaneously.

In this work, we demonstrate the capability of Continuous Wavelet Transform (CWT)-based representation (i.e., scalogram) for dysarthric severity-level classification. According to study in [5], wavelet transform has better frequency resolution in the low frequency regions, as compared to the STFT. In the literature, for acoustical research problems, wavelet-based features have been successfully implemented as in [3,22]. To that effect, the motivation of utilizing CWT for this study is the improved frequency resolution of CWT-based scalograms at lower frequencies as compared to the STFT-based and Mel spectrogram-based techniques. To the best of the authors' knowledge and belief, the use of CWT has been explored to Model Articulation Impairments in Patients with Parkinson's Disease [23]. However the use of CWT to capture discriminative acoustic cues for dysarthric severity-level classification is being proposed for the first time in this study. Results are presented on standard Universal Access (UA)-Speech Corpus.

The rest of paper is organized as follows: Sect. 2 discusses the motivation of using scalogram-based approach over a spectrogram. Section 3 describes the proposed approach of Morse wavelet-based dysarthric severity level classification. Furthermore, experimental setup is described in Sect. 4, followed by experimental results in Sect. 5. The Sect. 6 concludes the paper along with potential future research directions.

2 Spectrogram and Scalogram

STFT-based spectrograms are made up of windows of equal and fixed lengths that run across the length of the signal. As a result, in a spectrogram, the spread in time, as well as frequency-domains, remains constant throughout the time-frequency plane (i.e., constant time and frequency resolution). On the other hand, we can achieve *variable* time-frequency resolution by employing CWT-based representation (also known as scalogram). The time-frequency spread of the wavelet atoms $\psi_{u,s}$ determines the time-frequency resolution of scalogram. A Heisenberg box is defined by the spread in time multiplied by the spread in frequency in a time-frequency representation. In a scalogram, for low frequency regions, the spread in frequency is less, leading to a better frequency resolution, as shown by the boxes in Fig. 1. Furthermore, CWT can be computed using the wavelet $\psi_{u,s}(t)$, which has its Fourier transform denoted by $\hat{\psi}_{u,s}(\omega)$ [20]

Given that the center frequency of $\hat{\psi}(\omega)$ is indicated by η, the wavelet $\psi_{u,s}$ has a center frequency at $\frac{\eta}{s}$. The wavelet $\psi_{u,s}$ has an energy spread about the center frequency of $\psi_{u,s}$, which is given by [20]:

$$\frac{1}{2\pi}\int_{0}^{+\infty}\left(\omega-\frac{\eta}{s}\right)^{2}|\hat{\psi}_{u,s}(\omega)|^{2}d\omega=\frac{\sigma_{\omega}^{2}}{s^{2}}, \tag{1}$$

where,

$$\sigma_{\omega}^{2}=\frac{1}{2\pi}\int_{0}^{+\infty}(\omega-\eta)^{2}|\hat{\psi}(\omega)|^{2}d\omega. \tag{2}$$

Furthermore, the energy density in local time-frequency plane is denoted P_{Wf}, given by:

$$P_{W}f(u,\xi)=\left|Wf(u,s)\right|^{2}=\left|Wf\left(u,\frac{\eta}{\xi}\right)\right|^{2}. \tag{3}$$

The Eq. (3) is nothing but a scalogram with scaled time-frequency resolution.

Figure 1 shows the motivation behind choosing CWT-based approach over STFT. Energy conservation in STFT is [20]:

$$\int_{-\infty}^{+\infty}|f(t)|^{2}dt=\frac{1}{2\pi}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}|Sf(u,\zeta)|^{2}d\zeta du. \tag{4}$$

Energy conservation is preserved in analytic WT as well [20].

$$\int_{-\infty}^{+\infty} |f_a(t)|^2 dt = \frac{1}{C_\psi} \int_{0}^{+\infty} \int_{-\infty}^{+\infty} |Wf(u,s)|^2 du \frac{ds}{s^2}. \tag{5}$$

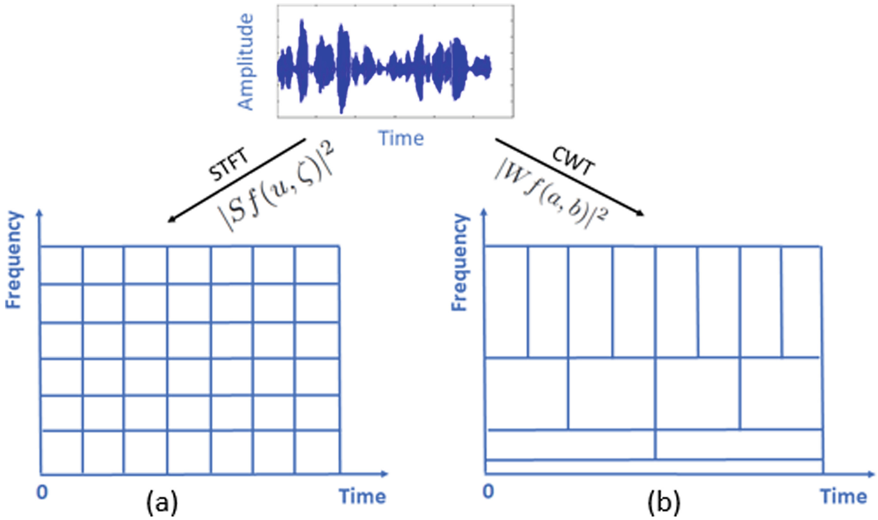

Fig. 1. Tilings of the time-frequency plane for (a) STFT and (b) CWT.

3 Proposed Work

3.1 Continuous Wavelet Transform (CWT)

Due to lack of co-ordination between brain and articulators, the speech produced by dysarthric patients have change in energy. To analyse this energy change in different severity-levels, recent investigations using spectrogram are made in [9]. However, to get better insight of energy spread in time-frequency representation, we propose CWT-based scalogram approach through this study. The key idea for employing CWT-based scalogram approach for dysarthric severity-level classification is to exploit the energy spread in the low frequency regions for different severity-levels in time-frequency distributions. A wavelet is a waveform with a zero-average and an effectively restricted duration, i.e., it is wave for a short duration and hence the name wavelet. It is defined as [17]:

$$\psi_{u,s}(t) = \frac{1}{\sqrt{s}} \psi^* \Big(\frac{t-u}{s} \Big), s \in R^+, u \in R, \tag{6}$$

where the dilation (scaling) parameter is denoted by s and the translational (positional) parameter is denoted by u. The CWT of a signal $f(t)$ is

$$\begin{aligned} W_f(u,s) &= <f(t), \psi_{u,s}(t)>, \\ &= \frac{1}{\sqrt{s}} \int_{-\infty}^{\infty} f(t) \psi^* \left(\frac{t-u}{s} \right) dt, \end{aligned} \tag{7}$$

where $<\cdot,\cdot>$ indicates inner product operation to compute wavelet coefficients, and $*$ denotes complex conjugate. The scalogram is defined as the square of absolute of the CWT coefficients, i.e., $|W_f(u,s)|^2$.

3.2 Exploiting Morse Wavelet for CWT

There are various types of analytic wavelets in the literature, such as Cauchy, complex Shannon, lognormal, Derivative of Gaussian, and Morlet wavelets [10, 13, 20]. However, due to the existence of various types of wavelets, choosing an appropriate wavelet for a particular task becomes an issue. Generalized Morse Wavelets (GMWs) is considered as a superfamily of analytic wavelets that are causal in the frequency-domain. In frequency-domain, the Morse wavelet is given by [16]:

$$\hat{\psi}_{\beta,\gamma}(\omega) = \int_{-\infty}^{\infty} \psi_{\beta,\gamma}(t) e^{-i\omega t} dt = U(\omega) a_{\beta,\gamma} \omega^{\beta} e^{-\omega^{\gamma}}, \tag{8}$$

where β and γ are the two parameters of the Morse wavelet, which control the *shape* and *size,* respectively, of the wavelet and $U(\omega)$ is unit-step function due to causality in the frequency-domain. The parameter β is called as the *order* and the parameter γ represents the *family* of wavelets. With each value of γ, one can get a family of wavelets from the Morse wavelet representation as shown in Eq. (8) [16]. The amplitude of the wavelet is normalized by a real-valued constant factor given by $\alpha_{\beta\gamma}$. The value of the constant scaling factor $\alpha_{\beta\gamma}$ is given by [17]:

$$\alpha_{\beta\gamma} \equiv 2 \left(\frac{e\gamma}{\beta} \right)^{\frac{\beta}{\gamma}}. \tag{9}$$

Furthermore, the "*wavelet duration*" denoted by $P^2_{\beta,\gamma}$ is given by the 2^{nd} order derivative of Morse wavelet. Mathematically, $P^2_{\beta,\gamma}$ can be defined as [17]:

$$P^2_{\beta,\gamma} \equiv - \frac{\omega^2_{\beta,\gamma} \hat{\psi}''_{\beta,\gamma}(\omega_{\beta,\gamma})}{\hat{\psi}_{\beta,\gamma}(\omega_{\beta,\gamma})} = \beta\gamma. \tag{10}$$

The number of peak frequency oscillations that may be fitted in the central window of a wavelet in the time-domain is given by $\frac{P^2_{\beta,\gamma}}{2\pi}$. The Morse wavelet with parameter $\gamma = 3$ (also known as *'Airy family'*) is used in this study. The optimum Heisenberg area $A_{\beta,\gamma}$, reached at $\gamma = 3$ even for a small wavelet duration (as shown in Fig. 2), justifies our choice of $\gamma = 3$. For a Morse wavelet, $A_{\beta,\gamma}$ is given by [5, 18]:

$$A_{\beta,\gamma} \equiv \sigma_t \sigma_\omega, \tag{11}$$

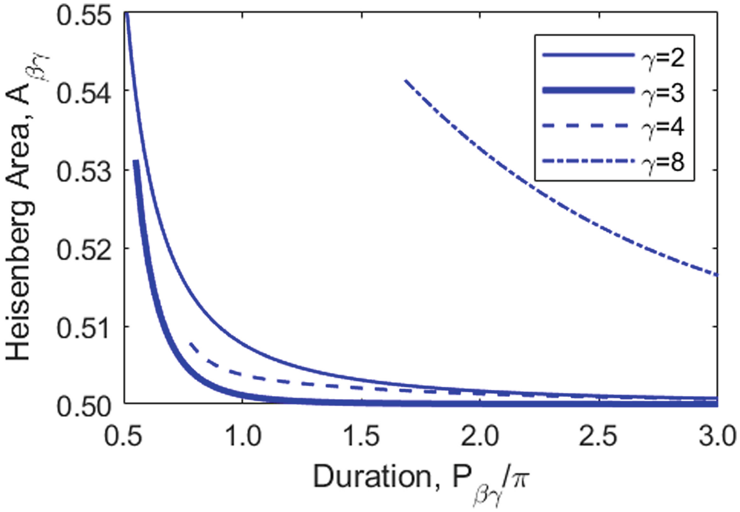

Fig. 2. Effect of γ parameter on the time-frequency Heisenberg area $A_{\beta,\gamma}$ w.r.t. wavelet duration $P_{\beta,\gamma}/\pi$. After [16].

where time spread σ_t^2 and frequency spread σ_ω^2 of wavelet atom representation are given by [17]:

$$\sigma_t^2 = \omega_\psi^2 \frac{\int t^2 |\psi(t)|^2 dt}{\int |\psi(t)|^2 dt} \quad \text{and,} \tag{12}$$

$$\sigma_\omega^2 = \frac{1}{\omega_\psi^2} \frac{\int (\omega - \tilde{\omega}_\psi)^2 |\psi(\omega)|^2 d\omega}{\int |\psi(\omega)|^2 d\omega}, \tag{13}$$

where $\tilde{\omega}_\psi$ represents the energy frequency of the Morse wavelet (which is also the mean of $|\Psi(\omega)|^2$) [17]. The study, reported in [16] shows that all the Morse wavelets attain the information concentration of $A_{\beta,\gamma} = 1/2$. For $\gamma = 3$, degree of concentration of information, i.e., $A_{\beta,\gamma}$ is the highest even for a small value of wavelet duration, $P_{\beta,\gamma}/\pi$, as shown in Fig. 2. To that effect, in this work, scalogram images were extracted using MATLAB with $\gamma = 3$ and $\beta = 20$ (i.e., $P_{\beta,\gamma}^2 = 60$) as the default parameter setting for Morse wavelet-based scalogram for full frequency band upto 8 kHz (since sampling frequency $F_s = 16$ kHz). Each scalogram image extracted is of $512 \times 512 \times 3$ dimension. These scalogram-based features are then fed as input to the CNN classifier. The experimental setup is explained in the following Section.

4 Experimental Setup

4.1 Dataset Used

The Universal Access dysarthric Speech (UA-Speech) corpus [25] is used to evaluate the proposed CWT-based approach. In this study, a dataset configuration

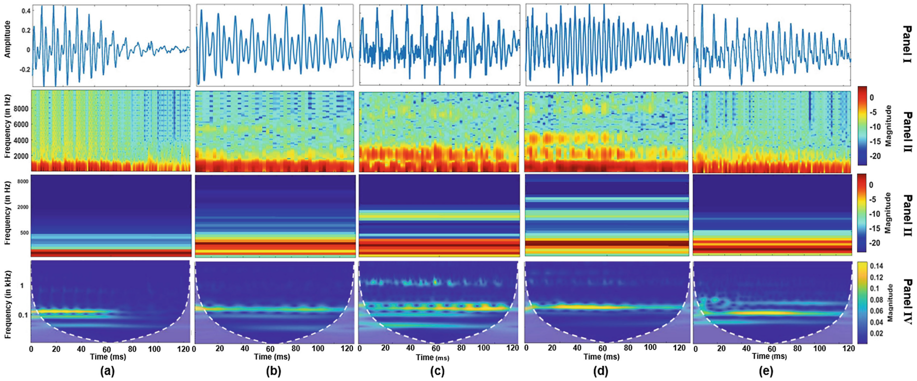

Fig. 3. Dysarthic speech utterance (for vowel /e/) for male speaker with various dysarthic severity-level (Panel I), corresponding STFT (Panel II), corresponding Mel spectrogram (Panel III), and corresponding Morse Wavelet Scalogram (Panel IV) for (a) normal, dysarthic speech with severity-level as (b) very low, (c) low, (d) medium, and (e) high. Best viewed in color. (Color figure online)

identical to that described in [9] is used. It has 8 speakers, out of which 4 are male and 4 are female speakers. Furthermore, 90% of the dataset is dedicated to training set and the remaining 10% is dedicated to the testing partition.

4.2 Feature Details

In this study, the energy capturing capabilities of scalogram at low frequencies are compared with the baseline spectrogram and Mel spectrogram. As mentioned in [9], the STFT was applied to generate a time-frequency representation with a window size of 2 ms, and window overlap of 0.5 ms. Furthermore, the performance of scalogram was also compared with Mel spectrogram, which are generated with a window of size 2 ms and overlap of 0.5 ms. The dimensions of the generated Mel spectrogram are 512 × 512 × 3. As discussed in Sect. 3, the scalograms of dimension 512 × 512 × 3 were generated with $\gamma = 3$, and $\beta = 20$ (i.e., $P_{\beta,\gamma} = 60$) as the default parameter setting.

4.3 Classifier Details

Based on the experiments presented in [12], the Convolutional Neural Network (CNN) is used as a classifier in this study. According to a study reported in [12], CNN gives comparable results with the other deep neural network (DNN)-based classifiers for the UA-Speech corpus. For this study, the CNN model was trained employing the Adam optimizer algorithm, four convolutional layers with kernel size of 5 × 5, and one Fully-Connected (FC) layer [14]. Mel spectrograms and scalograms, both of size 512 × 512, were used in these investigations. A max-pool layer and Rectified Linear Activation (ReLU) are utilised. For loss estimation, a learning rate of 0.001 and cross-entropy loss are chosen.

4.4 Performance Evaluation

F1-Score. It is a widely used statistical parameter for analyzing the performance of the model. As stated in [7], it is calculated as the harmonic mean of the model's precision and recall. Its value ranges from 0 to 1, with a score closer to 1 indicating higher performance.

MCC. It shows the degree of correlation between the expected and actual class [21]. For model comparison, it is typically regarded as a balanced measure. It is in the range of −1 to 1.

Jaccard Index. The Jaccard index is a metric for determining how similar and different the two classes are. It is in the range of 0 to 1. It is described as [2]:

$$\text{Jaccard Index} = \frac{TP}{TP + FP + FN}, \tag{14}$$

where TP, FP, and FN, represent True Positive, False Positive, and False Negative, respectively.

Hamming Loss. It considers class labels that were predicted wrongly. The prediction error (prediction of an incorrect label), and the missing error (prediction of a relevant label) are normalized across all the classes and test data. The following formula can be used to determine Hamming loss [6]:

$$\text{Hamming Loss} = \frac{1}{nL} \sum_{i=1}^{n} \sum_{j=1}^{L} I(y_i^j \neq \hat{y}_i^j), \tag{15}$$

where y_i^j and $\hat{y}_i^j$ are the actual and predicted labels, and I is an indicator function. The more it is close to 0, the better the performance of the algorithm.

5 Experimental Results

5.1 Spectrographic Analysis

Panel I of the Fig. 3 show the speech segment of vowel /e/. Panel II, III, and IV shows the spectrogram, Mel spectrogram, and scalogram, respectively, for (a) normal, (b) very low, (c) low, (d) medium, and (e) high dysarthric severity-level for the same speech segment. It can be observed from Fig. 4 that the scalogram-based features can capture energy-based discriminative acoustic cues for dysarthric severity-levels more accurately than the STFT and Mel spectrogram-based features. Furthermore, from scalogram, it can be observed that as the dysarthtic severity-level increases, patients struggle to speak the prolonged vowel, $/e/$. This may be due to the lack of coordination between articulators and the brain. Due to this, the energy spread is seen over the entire time-axis. However, the utterance of vowel $/e/$ is of short duration for medium and high dysarthtic severity-levels.

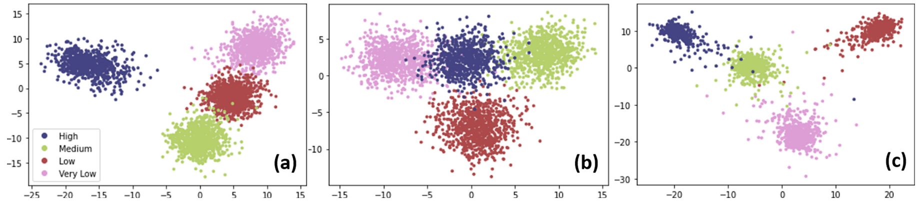

Fig. 4. Scatter plot obtained using LDA for (a) STFT, (b) Mel spectrogram, and (c) Scalogram. After [11]. Best viewed in color.

Table 1. Results in (% classification accuracy) for CNN classifier.

Feature set	CNN
STFT	91.76
Mel-spectrogram	92.65
Scalogram	**95.17**

5.2 Performance Evaluation

The performance evaluation for various feature sets is done *via* % classification accuracy (as shown in Table 1). On CNN, the scalogram performs relatively better with a classification accuracy of 95.17% than the baseline STFT, and Mel spectrogram. The analyses in the following sub-Section and the % classification accuracy obtained through the CNN classifier, show the capabilities of the scalogram in capturing the energy spread generated during the speech production mechanism for various dysarthric severity-level. Furthermore, Table 2 shows the confusion matrix of the STFT, Mel spectrogram, and scalogram for CNN model. It can be observed that the scalogram reduces the false prediction error, which indicates the better performance of the scalogram $w.r.t$ the baseline STFT, and Mel spectrogram. Additionally, Table 3 shows the comparison between statistical measures using the F-1 score, Jaccard index, MCC, and Hamming loss for various feature sets. It can be observed from Table 3 that scalogram performs relatively better than the baseline STFT and Mel spectrogram.

5.3 Visualization of Various Features Using Linear Discriminant Analysis (LDA)

The capabilities of scalogram for the classification of the dysarthic severity-level is also validated by LDA scatter plots due to it's higher image resolution and better projection of the given higher-dimensional feature space to lower-dimensional than the scatter plots obtained using t-sne plots [11]. Here, the LDA plot of STFT, Mel spectrogram, and scalogram are projected onto 2-D feature space, and represented using the scatter plot shown in Fig. 4 (a), Fig. 4 (b), and Fig. 4 (c), respectively. From Fig. 4, it can be observed that wavelet-based scalogram

Table 2. Confusion matrix obtained for STFT, Mel-spectrogram, and scalogram.

Feature Set	Severity	High	Medium	Low	Very Low
STFT	High	63	6	3	3
	Medium	10	79	3	1
	Low	3	4	79	7
	Very Low	1	2	1	89
Mel-Spectrogram	High	69	1	3	2
	Medium	5	81	4	3
	Low	4	1	91	0
	Very Low	4	0	2	89
Scalogram (Morse Wavelet)	High	69	5	1	0
	Medium	3	89	1	0
	Low	1	1	90	1
	Very Low	3	0	1	89

Table 3. Various statistical measures of STFT, Mel spectrogram, and scalogram.

Feature set	F1-score	MCC	Jaccard index	Hamming loss
STFT	0.87	0.83	0.776	0.124
Mel spectrogram	0.92	0.90	0.86	0.073
Scalogram	**0.95**	**0.91**	**0.94**	**0.05**

has low intra-class variance and high inter-class variance, which increases the distance between the clusters *w.r.t* baseline STFT, and Mel spectrogram, thereby better classification performance by the Morse wavelet.

6 Summary and Conclusion

In this study, we investigated CWT, in particular, the Morse wavelet, to achieve improved resolution in time and frequency representation for various dysarthric severity levels. The low-frequency resolution of Morse wavelet-based scalogram is higher than the resolution of STFT and Mel spectrogram. Therefore, the energy spread corresponding to the dysarthric severity in low-frequency region is better visualized in the scalogram. Hence, the low-frequency discriminative cues are better classified using a scalogram. This can also be observed with the significant increase in % classification accuracy as compared to the STFT and Mel spectrogram. Furthermore, it was also observed that as the severity-level increases, due to difficulty for patients to utter the complete word, the energy spreading is more in frequency representation over the entire time-axis. The performance of the scalogram is also analyzed using various statistical performance parameters, such as $F1$-Score, MCC, Jaccard Index, Hamming Loss, and LDA scatter plots. Other dysarthric speech corpora, such as TORGO and Homeservice, will be used to further validate this work in the future. Our future efforts will focus on extending and validating this work on other dysarthric speech corpora, such as TORGO and Home service.

Acknowledgments. The authors would like to thank the Ministry of Electronics and Information Technology (MeitY), New Delhi, Govt. of India, for sponsoring the consortium project titled 'Speech Technologies in Indian Languages' under 'National Language Translation Mission (NLTM): BHASHINI', subtitled 'Building Assistive Speech Technologies for the Challenged' (Grant ID: 11(1)2022-HCC (TDIL)). We also thank the consortium leaders Prof. Hema A. Murthy, Prof. S. Umesh of IIT Madras, and the authorities of DA-IICT Gandhinagar, India for their support and cooperation to carry out this research work.

Appendix

A.1. Energy Conservation in STFT

The energy conservation in STFT for any signal $f(t) \in L^2(R)$ is given by

$$\int_{-\infty}^{+\infty} |f(t)|^2 dt = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} |Sf(u,\zeta)|^2 d\zeta du, \tag{16}$$

Here, u and ζ indicate the time-frequency indices that vary across R and hence, covers the entire time-frequency plane. The reconstruction of signal can then be given by

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} Sf(u,\zeta) g(t-u) e^{i\zeta t} d\zeta du. \tag{17}$$

Applying Parseval's formula to Eq. (17) w.r.t. to the integration in u, we get

$$Sf(u,\zeta) = e^{-iu\zeta} f * g_\zeta(u), \tag{18}$$

Here, $g_\zeta(t) = g(t)e^{i\zeta t}$. Hence, Fourier Transform of $Sf(u,\zeta)$ is $\hat{f}(\omega_\zeta)\hat{g}(\omega)$. Furthermore, after applying the Plancherel's formula to Eq. (16) gives

$$\frac{1}{2\pi} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} |Sf(u,\zeta)|^2 du d\zeta = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \frac{1}{2\pi} \int_{-\infty}^{+\infty} |\hat{f}(\omega+\zeta)\hat{g}(\omega)|^2 d\omega d\zeta. \tag{19}$$

Finally, the Plancheral formula and the Fubini theorem result in $\frac{1}{2\pi} \int_{-\infty}^{+\infty} |\hat{f}(\omega+\zeta)|^2 d\zeta = ||f||^2$, which validates STFT's energy conservation as demonstrated in Eq. (16), It explains why the overall signal energy is the same as the time-frequency sum of the STFT.

A.2. Energy Conservation in CWT

Using the same derivations as in the discussion of Eq. 17, one can verify that the inverse wavelet formula reconstructs the analytic part of f :

$$f_a(t) = \frac{1}{C_\psi}\int_0^{+\infty}\int_{-\infty}^{+\infty} Wf_a(u,s)\psi_s(t-u)\frac{ds}{s^2}du. \tag{20}$$

Applying the Plancherel formula for energy conservation for the analytic part of f_a given by

$$\int_{-\infty}^{+\infty}|f_a(t)|^2dt = \frac{1}{C_\psi}\int_0^{+\infty}\int_{-\infty}^{+\infty}|W_af(u,s)|^2du\frac{ds}{s^2}. \tag{21}$$

Since $Wf_a(u,s) = 2Wf(u,s)$ and $||f_a||^2 = 2||f||^2$. If f is real, and the variable change $\zeta = \frac{1}{s}$ in energy conservation denotes that

$$||f||^2 = \frac{2}{C_\psi}\int_0^{+\infty}\int_{-\infty}^{+\infty}P_wf(u,\zeta)dud\zeta. \tag{22}$$

It reinforces the notion that a scalogram represents a time-frequency energy density.

References

1. Al-Qatab, B.A., Mustafa, M.B.: Classification of dysarthric speech according to the severity of impairment: an analysis of acoustic features. IEEE Access **9**, 18183–18194 (2021)
2. Bouchard, M., Jousselme, A.L., Doré, P.E.: A proof for the positive definiteness of the Jaccard index matrix. Int. J. Approx. Reason. **54**(5), 615–626 (2013)
3. Chen, H., Zhang, P., Bai, H., Yuan, Q., Bao, X., Yan, Y.: Deep convolutional neural network with scalogram for audio scene modeling. In: INTERSPEECH, Hyderabad India, pp. 3304–3308 (2018)
4. Darley, F.L., Aronson, A.E., Brown, J.R.: Differential diagnostic patterns of dysarthria. J. Speech Hear. Res. (JSLHR) **12**(2), 246–269 (1969)
5. Daubechies, I.: The wavelet transform, time-frequency localization and signal analysis. IEEE Trans. Inf. Theory **36**(5), 961–1005 (1990)
6. Dembczyński, K., Waegeman, W., Cheng, W., Hüllermeier, E.: Regret analysis for performance metrics in multi-label classification: the case of hamming and subset zero-one loss. In: Balcázar, J.L., Bonchi, F., Gionis, A., Sebag, M. (eds.) ECML PKDD 2010. LNCS (LNAI), vol. 6321, pp. 280–295. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15880-3_24
7. Fawcett, T.: An introduction to ROC analysis. Pattern Recognit. Lett. **27**(8), 861–874 (2006)
8. Gillespie, S., Logan, Y.Y., Moore, E., Laures-Gore, J., Russell, S., Patel, R.: Cross-database models for the classification of dysarthria presence. In: INTERSPEECH, Stockholm, Sweden, pp. 3127–31 (2017)
9. Gupta et al., S.: Residual neural network precisely quantifies dysarthria severity-level based on short-duration speech segments. Neural Netw. **139**, 105–117 (2021)
10. Holschneider, M.: Wavelets. An analysis tool (1995)

11. Izenman, A.J.: Linear discriminant analysis. In: Izenman, A.J. (ed.) Modern Multivariate Statistical Techniques. Springer Texts in Statistics, pp. 237–280. Springer, New York (2013). https://doi.org/10.1007/978-0-387-78189-1_8
12. Joshy, A.A., Rajan, R.: Automated dysarthria severity classification using deep learning frameworks. In: 28th European Signal Processing Conference (EUSIPCO), Amsterdam, Netherlands, pp. 116–120 (2021)
13. Knutsson, H., Westin, C.F., Granlund, G.: Local multiscale frequency and bandwidth estimation. In: Proceedings of 1st International Conference on Image Processing, Austin, TX, USA, vol. 1, pp. 36–40, 13–16 November 1994
14. LeCun, Y., Kavukcuoglu, K., Farabet, C.: Convolutional networks and applications in vision. In: Proceedings of 2010 IEEE International Symposium on Circuits and Systems, Paris, France, pp. 253–256 (2010)
15. Lieberman, P.: Primate vocalizations and human linguistic ability. J. Acoust. Soci. Am. (JASA) **44**(6), 1574–1584 (1968)
16. Lilly, J.M., Olhede, S.C.: Generalized Morse wavelets as a superfamily of analytic wavelets. IEEE Trans. Signal Process. **60**(11), 6036–6041 (2012)
17. Lilly, J.M., Olhede, S.C.: Higher-order properties of analytic wavelets. IEEE Trans. Signal Process. **57**(1), 146–160 (2008)
18. Lilly, J.M., Olhede, S.C.: On the analytic wavelet transform. IEEE Trans. Inf. Theory **56**(8), 4135–4156 (2010)
19. Mackenzie, C., Lowit, A.: Behavioural intervention effects in dysarthria following stroke: communication effectiveness, intelligibility and dysarthria impact. Int. J. Lang. Commun. Disord. **42**(2), 131–153 (2007)
20. Mallat, S.: A Wavelet Tour of Signal Processing, 2nd edn. Elsevier, Amsterdam (1999)
21. Matthews, B.W.: Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA) Prot. Struct. **405**(2), 442–451 (1975)
22. Ren, Z., Qian, K., Zhang, Z., Pandit, V., Baird, A., Schuller, B.: Deep scalogram representations for acoustic scene classification. IEEE/CAA J. Automatica Sinica **5**(3), 662–669 (2018)
23. Vásquez-Correa, J.C., Orozco-Arroyave, J.R., Nöth, E.: Convolutional neural network to model articulation impairments in patients with Parkinson's disease. In: INTERSPEECH, Stockholm, pp. 314–318 (2017)
24. Young, V., Mihailidis, A.: Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review. Assist. Technol. **22**(2), 99–112 (2010)
25. Yu, J., et al.: Development of the CUHK dysarthric speech recognition system for the UA speech corpus. In: INTERSPEECH, Hyderabad, India, pp. 2938–2942 (2018)