Neural Networks 139 (2021) 105–117

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Residual Neural Network precisely quantifies dysarthria severity-level based on short-duration speech segments

Siddhant Gupta [a], Ankur T. Patil [a], Mirali Purohit [a], Mihir Parmar [b], Maitreya Patel [a], Hemant A. Patil [a], Rodrigo Capobianco Guido [c,*]

[a] Speech Research Lab, Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar 382007, India
[b] Arizona State University, Tempe, USA
[c] Instituto de Biociências, Letras e Ciências Exatas, Unesp - Univ Estadual Paulista (São Paulo State University), Rua Cristóvão Colombo 2265, Jd Nazareth, 15054-000, São José do Rio Preto - SP, Brazil

ARTICLE INFO

Article history:
Available online 24 February 2021

Keywords:
Dysarthria
Severity-level
Short-speech segments
CNN
ResNet

ABSTRACT

Recently, we have witnessed Deep Learning methodologies gaining significant attention for severity-based classification of dysarthric speech. Detecting dysarthria, quantifying its severity, are of paramount importance in various real-life applications, such as the assessment of patients' progression in treatments, which includes an adequate planning of their therapy and the improvement of speech-based interactive systems in order to handle pathologically-affected voices automatically. Notably, current speech-powered tools often deal with short-duration speech segments and, consequently, are less efficient in dealing with impaired speech, even by using Convolutional Neural Networks (CNNs). Thus, detecting dysarthria severity-level based on short speech segments might help in improving the performance and applicability of those systems. To achieve this goal, we propose a novel Residual Network (ResNet)-based technique which receives short-duration speech segments as input. Statistically meaningful objective analysis of our experiments, reported over standard Universal Access corpus, exhibits average values of 21.35% and 22.48% improvement, compared to the baseline CNN, in terms of classification accuracy and F1-score, respectively. For additional comparisons, tests with Gaussian Mixture Models and Light CNNs were also performed. Overall, the values of 98.90% and 98.00% for classification accuracy and F1-score, respectively, were obtained with the proposed ResNet approach, confirming its efficacy and reassuring its practical applicability.

1. Introduction

Dysarthria (Freed, 2018) consists of a motor speech disorder in which the articulatory elements and muscles required to speak ordinarily are somehow affected, paralyzed or damaged. Individuals suffering from dysarthria face difficulties in conveying a spoken message or expressing voice emotions, since vocal folds, tongue, and associated muscles cannot be adequately controlled.

Concomitantly with dysarthria-related issues, we know that, with the advancement of speech technologies, Intelligent Personal Assistants (IPAs) such as *Google Assistant*, *Siri*, *Amazon*, and *Alexa*, are overgrowing. Nevertheless, these systems have been produced based on the assumption that the speech to be processed is in its natural form and comes from a healthy subject. Hence, they are not capable of performing speech recognition efficiently in impaired people (Young & Mihailidis, 2010). Moreover, recent interactive devices, such as specific IPAs, use to deal with short-duration speech segments and, consequently, their performance is highly dependent on the Automatic Speech Recognition (ASR) algorithm (De Russis & Corno, 2019; Swarup, Maas, Garimella, Mallidi, & Hoffmeister, 2019).

Not only ASR systems benefit from the possibility of assessing and identifying dysarthric speech (Bhat, Vachhani, & Kopparapu, 2017a; Mustafa, Salim, Mohamed, Al-Qatab, & Siong, 2014) but also many other systems. To allow for dysarthric speech enhancement and patients' progression in treatment, detecting the severity-level of a pathology from short-duration speech segments is an essential task. Standard methods for the assessment of dysarthric speech are traditionally based on a clinical trial by Speech Language Pathologists (SLPs), using pre-defined rating scales or observing the movement of various articulatory elements over the spoken time-interval (Rudzicz, Namasivayam, & Wolff, 2012). For dysarthria detection, specific speech segments from a certain set of speakers can be obtained from long speech signals based on manual or automatic framing. In the latter case, deep learning-based approaches have played an important role,

* Corresponding author.
E-mail address: guido@ieee.org (R.C. Guido).

https://doi.org/10.1016/j.neunet.2021.02.008
0893-6080/

as demonstrated in papers (Wang & Chen, 2018; Zhang & Wu, 2013), and (Zhang & Wang, 2016).

The recent study documented in Connaghan, Wertheim, Laures-Gore, Russell, and Patel (2020) demonstrates how important semantic differential scales are to investigate listener impressions on speakers with dysarthria. For the severity-based classification, most of the methods focus on feature-based techniques and acoustic modeling (Calvert, Spence, Stein, et al., 2004; Falk, Chan, & Shein, 2012; Paja & Falk, 2012a; Sztahó & Vicsi, 2016). Accordingly, the study reported in Gurevich and Scamihorn (2017), for instance, shows a detailed analysis of the various methodologies for dysarthria severity assessment using different rating scales defined decades ago (Enderby, 1980; Fahn & Elton, 2000; Hoehn & Yahr, 1967; Schmitz-Hübsch et al., 2006; Yorkston, Beukelman, & Traynor, 1984a, 1984b). There is, however, substantial performance variability among listeners who transcribe degraded speech, as mentioned by the authors of paper (kyong Choe, Liss, Azuma, & Mathy, 2012). The authors of paper (Paja & Falk, 2012b) proposed the application of a Mahalanobis distance-based discriminant classifier in conjunction with a set of acoustic features formerly proposed for intelligibility prediction and voice pathology assessment, where feature selection is used to sieve features for both disorder severity classification and intelligibility prediction. In the same way, multinomial logistic regression with sparsity constraints is used in paper (Lansford, Berisha, & Utianski, 2016) as a similarity-based approach to characterize dysarthria. In sum, however, those processes are time-consuming, difficult, and costly.

Turning to more recent computer-based assessments, as reported in paper (Gomez et al., 2019), for instance, 16 subjects affected with Parkinson disease and 16 healthy subjects have shown considerable differences between the statistical distributions of dynamic articulation features based on Kullback–Leibler and Jensen–Shannon divergences. Accordingly and as explained in paper (Yang et al., 2020), a total of 35 patients affected by Parkinson disease and 26 healthy controls were considered to perform single-, double-, and multiple-syllable tests based on logistic regression. The corresponding results revealed that the minimum, maximum and mean fundamental frequencies, in addition to jitter, duration of speech, and median intensity of speaking were considerably different for the groups. In the same direction, the authors of paper (Giri & Rayavarapu, 2018) demonstrated a relatively similar result.

Some artificial intelligence-based approaches have been tried, including Hidden Markov Models (HMMs) using a Maximum Likelihood Estimation (MLE) technique (Bhat, Das, Vachhani, & Kopparapu, 2018; Bhat, Vachhani, & Kopparapu, 2016), Gaussian Mixture Models (GMMs), and Long Short-Term Memory (LSTM) (Kim, Cao, An, & Wang, 2018). Another similar strategy was proposed in Bhat et al. (2017a): it uses different feature sets combined with Artificial Neural Networks (ANNs), getting worth-mentioning results in severity-based classification. Data augmentation technique was also used to address the data scarcity problem in severity-based classification (Vachhani, Bhat, & Kopparapu, 2018). Following the same way, the authors of paper (Bhat, Vachhani, & Kopparapu, 2017b) proposed a non-linguistic manner for the automatic assessment of severity levels of dysarthria by means of music-related features. An ordinary Artificial Neural Network (ANN) was used together with Universal Access (UA) corpus and TORGO database, where the average classification values of accuracy of 96.44% and 98.7% were obtained for those datasets, respectively. Nevertheless, all the above-mentioned strategies are sub-optimal in particular aspects, with limitations in at least one specific sense, such as being based solely in shallow classifiers.

Proceeding over time, we can find another significant piece of work in paper (Chandrashekar, Karjigi, & Sreedevi, 2019), which combined spectro-temporal features, ANNs, and Convolutional Neural Network (CNN)-based classifiers. Accordingly, the results reported in paper (Farhadipour, Veisi, Asgari, & Keyvanrad, 2018), in which authors presented a feature-extraction method based on Deep Belief Networks (DBNs) to identify speakers suffering from dysarthria, are relevant: considering UA corpus, an accuracy of 97.3% was reported. Following, Vásquez Correa, Arias, Orozco-Arroyave and Nöth (2018) the authors proposed a multitask learning approach based on CNNs to discover different patients' speech issues, such as lack of possibility to move the lips, larynx, tongue and palate. They found that their approach improved in 4% the average accuracy in relation to single networks trained to analyze and evaluate speech details.

Another interesting piece of work published by the authors of paper (Perez, Aldeneh, & Provost, 2020) consists of a novel acoustic model based on a mixture of experts, allowing for intelligibility speech stages to be assessed. The proposed approach drastically reduced phone error rates across all severity stages in aphasic speech, in comparison with their baseline strategy. In the same way, the authors of paper (An et al., 2018) investigated the possible existence of amyotrophic lateral sclerosis based on dysarthria. Different CNN structures were used, considering both time-domain and frequency-domain. Experimental results outperformed ordinary ANN-based approaches, with results around 71.6% and 80.9% for the values of sensitivity and specificity, respectively. Similarly, the authors of paper (Tripathi, Bhosale, & Kopparapu, 2020) proposed a speaker-independent intelligibility assessment system based on DeepSpeech, which is an end-to-end speech-to-text engine, and a support vector machine. Considering the UA corpus, a value of accuracy of 53.9% was obtained.

Even adopting deep classifiers, all the previous approaches proposed for dysarthria severity classification work effectively on long-duration speech signals, i.e., those generally ranging from 4 to 8 s. In this study, however, a novel approach has been proposed for severity-based classification of dysarthric speech based on short-duration speech segments, lasting less than one second. In addition, inspired on two recently-proposed CNN-based approaches used to detect Parkinson's disease, as documented in papers (Vásquez Correa, Arias-Vergara, Orozco-Arroyave and Nöth, 2018; Vásquez-Correa, Orozco-Arroyave, & Nöth, 2017), our strategy focuses on a Residual Network (ResNet)-based classification algorithm.

ResNet was introduced in 2015 and revolutionized the field of Deep Neural Networks as it was able to achieve a network depth of more than 100 layers, which was far deeper than the other existing neural network architectures of those times. Since then, ResNet has proved to be very successful in image classification (Jiang, Chen, Zhang, & Xiao, 2019; Liu, Tian, & Xu, 2019; Wang et al., 2017), image recognition (He, Zhang, Ren, & Sun, 2016a; Lu, Jiang, & Kot, 2018), and computer vision applications (Jung et al., 2017; Liu et al., 2020). Inspired by these successes, the effectiveness ResNet model has also been explored for the speech research problems. In Chen, Xie, Zhang, and Xu (2017), ResNet is used to classify between the speech of a genuine speaker and replayed speech on ASVspoof2017 dataset. The model was able to achieve an Equal Error Rate of 16.26 when MFCC features were used as input to the model. The authors of Vydana and Vuppala (2017) used a Hidden Markov Model-based ResNet for speech recognition task. While comparing the results with a Hidden Markov Model-based Deep Neural Network, a reduction in Word Error Rate of 8% was achieved. However, to the best of author's knowledge, ResNet architecture has not been used for classification of dysarthric speech and hence, this is the first time that the ResNet architecture has been employed for severity-level classification of dysarthric speech.

Therefore, based on the limitations of key methods, previously reviewed, the proposed approach successfully advances

Table 1
Severity classification based on intelligibility.
Source: Adapted from Kim et al. (2008).

Intelligibility rating (%)	Severity-level
0–25	High
25–50	Medium
50–75	Low
75–100	Very low

the state-of-the-art in the field. Our specific contributions are summarized as follows: (i) this is the first attempt of its kind to detect the severity of dysarthric speech using short-duration speech segment; (ii) this is the first time ResNet is adopted for dysarthria severity classification; and (iii) statistically meaningful experiments on standard UA corpus were conducted, reassuring the efficacy of our original technique.

Aiming at a better understanding of the concepts explained hereafter, the remaining of this paper is structured as follows: Section 2 presents the problem statement, Section 3 characterizes dysarthric speech to investigate suitable feature representation and nonlinearities in speech production. Section 4 explains the proposed methodology and Section 5 shows the experimental setup for which the corresponding results are discussed in its subsections. Lastly, Section 6 describes the conclusions along with future suggested directions.

2. Problem formulation

Our goal is to classify dysarthric speech based on its severity-level using short-duration speech segment. In this study, dysarthric speech is classified into four severity-based categories as shown in Table 1. As suggested in Kim et al. (2008), five naive listeners were recruited for each speaker, and they were allowed to listen to words as many times as needed for transcription. For each listener's transcription, the percentage of correct responses was calculated. The correct percentage was then averaged across five listeners to obtain each speaker's intelligibility. Based on the averaged percent accuracy, each speaker was classified into one of four categories as shown in Table 1.

Since speech is essentially produced upon air exhalation, a precisely coordinated respiratory support is of paramount importance for communication. In dysarthric subjects, however, the combined pneumo-phono-articulatory cognitive commands from the brain are pathologically-affected, drastically degrading speech quality. Remarkably observable, distorted vowel sounds have been a direct consequence of dysarthria, where articulatory undershoot forces a humble vowel working space, as mentioned in papers (Lansford & Liss, 2014; de Oliveira Chappaz, dos Santos Barreto, & Ortiz, 2018). Hence, as shown in papers (Kim, Hasegawa-Johnson and Perlman, 2011; Kim, Kent and Weismer, 2011; Rosen, Goozee, & Murdoch, 2008; Turner, Tjaden, & Weismer, 1995; Watanabe, Arasaki, Nagata, & Shouji, 1994), formant frequencies centralization, uncommon formant frequencies for both front and high vowels, formants instability, and reduced slopes involving the second formant, are notable. This justifies our efforts in using short speech segments for the detection of dysarthria, since, presumably, they contain the formant-related information we need and, in addition, are capable of characterizing severity-levels. Based on our strong evidences, let us move forward to the formal problem formulation.

Let $\mathcal{X} = \{x_i\}_{i=1}^n$ denote the features of dysarthric speech, and $\mathcal{Y} = \{y_i\}_{i=1}^n$ denotes the corresponding labels. First, we map this labeled data, $\mathcal{D} = \{x_i, y_i\}_{i=1}^n$ as:

$$f(x) = \begin{cases} y = 0, & \text{if severity-level is high,} \\ y = 1, & \text{if severity-level is mid,} \\ y = 2, & \text{if severity-level is low,} \\ y = 3, & \text{if severity-level is very low.} \end{cases}$$

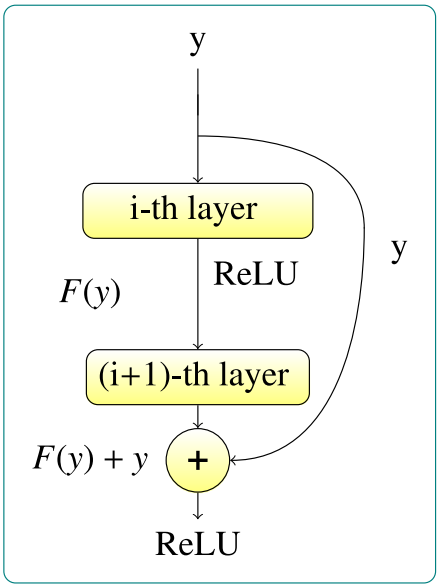


Fig. 1. The residual block strategy.

Problem Statement: Given a manually annotated dysarthric speech data ($\mathcal{D}$), for severity-based classification in four categories, learn severity-based classifier (as a mapping function), $\mathcal{F} : \mathcal{X} \rightarrow \mathcal{Y}$, which can do efficient classification using short-duration speech.

To solve our proposed problem, we certainly need an adequate classifier. Although universal approximation theory (Christensen & Christensen, 2006) presents results allowing for the conclusion that feedforward neural networks containing a single layer could represent any function, data overfitting and the vanishing gradient issue have forced machine learning algorithms to advance much more. As observed in practice and confirmed theoretically, however, expanding the networks in such a way they get deeper does not mean just adding layers because accuracy and performance might degrade extraordinarily fast. Thus, since they allow for training up to thousands of layers with remarkable performance, Deep Residual Networks (ResNets) (Kawaguchi & Bengio, 2019) have been considered one of the most groundbreaking advancements in deep neural network-related fields.

The *identity shortcut connections* (ISCs), used to occasionally skip one or more network layers, as shown in Fig. 1, is the essence of ResNets. By blocking the information flow in case the feature maps of two subsequent layers present considerably different distributions, usually jumping two or three layers, ISCs allow for clear and appropriate gradient paths. Simulating the pyramidal cells in the cerebral cortex, the circumvention permits background and rectifier linear unit (ReLU) normalization (Huang, Liu, Weinberger, & van der Maaten, 2017). Therefore, since ResNets have been the most flexible structure capable of handling the above-mentioned problems adequately, they are presumably useful to capture the differences between healthy and dysarthric subjects. Furthermore, considering that they had never been used to the detection of dysarthria, we are adopting them as being our fundamental classification strategy.

3. Characterizing dysarthria in speech signals

In this section, we present the time-domain, frequency-domain, and joint time–frequency domain analysis of dysarthric

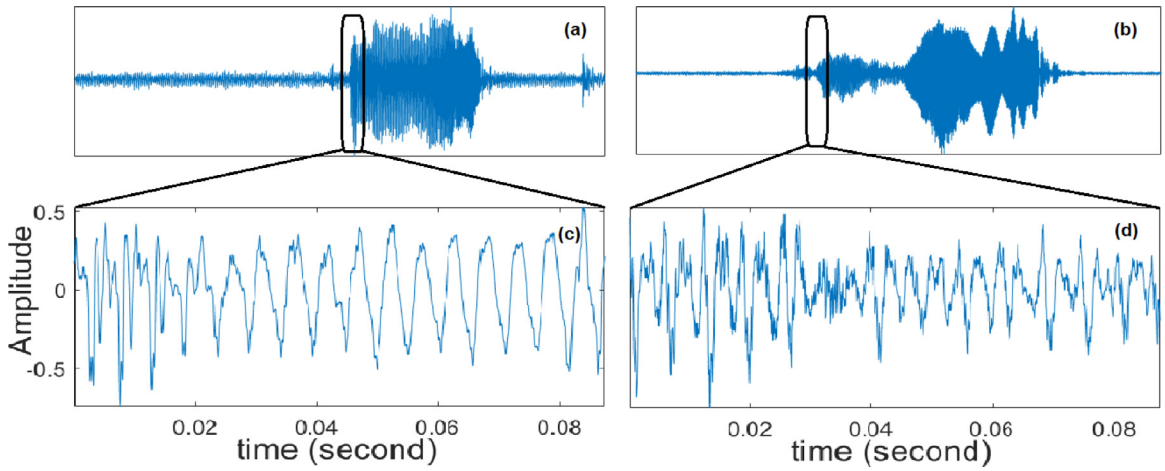


Fig. 2. Analysis of time-domain waveform: (c) is the segment of normal speech signal which is depicted in (a), and (d) is the segment of the dysarthric speech signal which is depicted in (b).

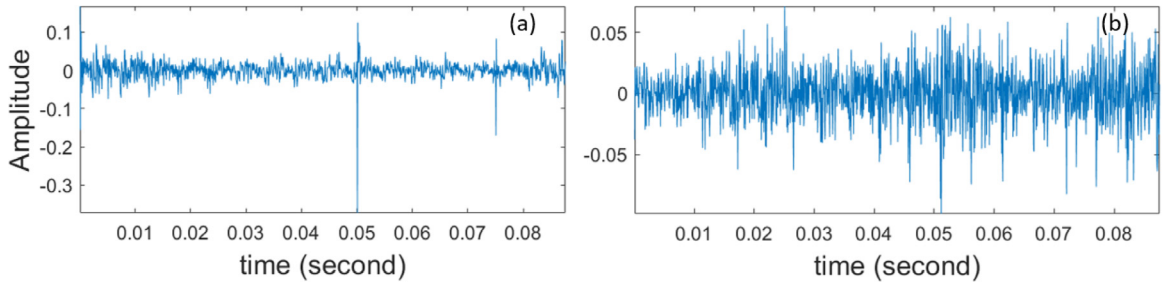


Fig. 3. Analysis via LP residual for signals shown in Fig. 2(a) LP residual for normal speech, and (b) LP residual for dysarthric speech.

speech as compared to its normal counterpart. The key motivation for such an analysis is to be able to choose appropriate feature representation for proposed deep learning architecture in our paper.

3.1. Analysis of time-domain waveform and glottal excitation source

Fig. 2-(a) and (b) show the time-domain waveform for normal and dysarthric speech, respectively. It can be observed from time-domain acoustic pressure variations that dysarthric speech has several distinct characteristics, such as relatively longer duration, longer pitch period, and production noise, possibly due to frication. This is a direct consequence of the combined pneumo-phono-articulatory cognitive commands from pathologically-affected brains, as discussed in the previous section.

In order to analyze the characteristics of speech excitation source during production, we employed two well-known methods, namely, Linear Prediction (LP) residual and Teager Energy Operator (TEO) profile. In particular, for a speech signal $s(n)$, LP residual is given by (Atal & Hanauer, 1971)

$$r(n) = s(n) - \bar{s}(n),$$

where $\bar{s}(n) = \sum_{k=1}^{p} a_k \cdot s(n-k)$, with a_k representing the Linear Prediction Coefficients (LPC). LP residual is known to give relatively higher values at, and also around, Glottal Closure Instants (GCIs) primarily due to relatively weaker capability of linear predictor before excitation signal is applied to, or reaches, the vocal tract system, as it has its own inertia so far as physics of speech production is concerned. Hence, historically LP residual and its analytic representation via Hilbert transform are used for estimation of GCIs (Ananthapadmanabha & Yegnanarayana, 1975, 1979). To that effect, Fig. 3 shows the plot of LP residual for normal vs. dysarthic speech cases shown in Fig. 2-(a) and (b). We can observe from Fig. 3 that LP residual is highly irregular for dysarthic speech, indicating abnormal changes in pitch period (T_0) and hence, pitch frequency (or fundamental frequency, F_0). That is why changes in T_0, i.e., jitter, have been used in literature for classification of normal vs. dysarthric speech.

Next, we present the analysis of the excitation source part via Teager Energy Operator (TEO) profile. In particular, for a speech signal $s(n)$, TEO profile is given by

$$\text{TEO}\{s(n)\} = (s(n))^2 - s(n-1) \cdot s(n+1) \quad .$$

From TEO, we can observe that three consecutive speech samples are required to find the running estimate of signal energy and, thus, it is known to have excellent time-resolution (Kaiser, 1990). Fig. 4 shows the corresponding TEO profile for the normal vs. dysarthic speech case shown in Fig. 2. We can note from Fig. 4 that, as in LP residual, TEO is also highly irregular for dysarthic speech, indicating abnormal changes in pitch period, i.e., T_0, and, hence, pitch frequency. In particular, TEO is found to give high energy pulses corresponding to GCIs due to its capability to capture characteristics of impulse-like excitation which are known to have higher signal-to-noise (SNR) ratios.

3.2. Analysis of nonlinearities in speech production

Historically, TEO was developed to investigate possible nonlinearities in speech production, showing bumps within glottal cycles. If production mechanisms for speech would have been linear, then the corresponding impulse response of each 2nd order digital resonator, whose cascade approximates frequency response of vocal tract system, would have been as damped sinusoids and, hence, corresponding TEO profile would be an exponentially decaying function (Kaiser, 1990).

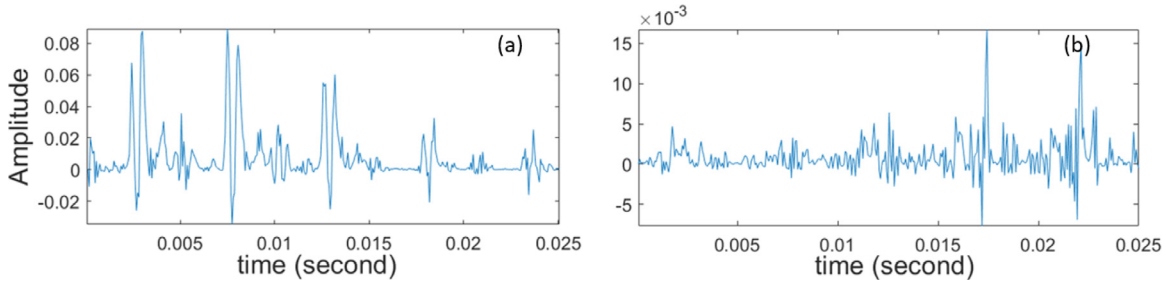


Fig. 4. Analysis via TEO profile for signals shown in Fig. 2, as follows: (a) TEO profile for normal speech, and (b) TEO profile for dysarthric speech.

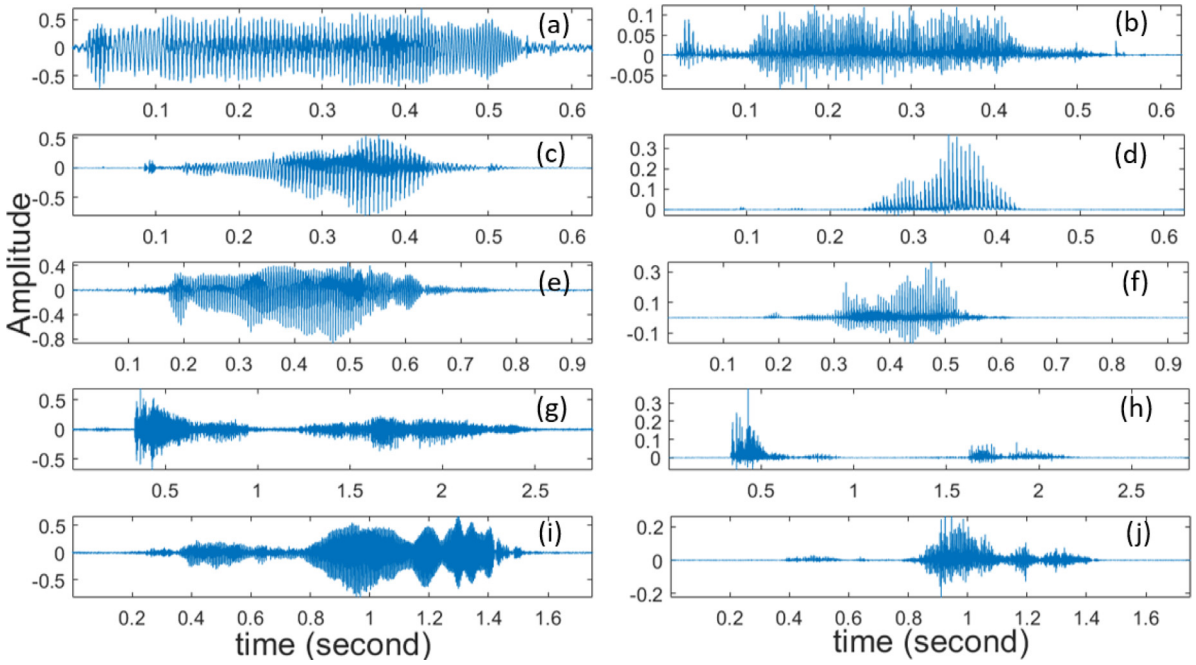


Fig. 5. Analysis of nonlinearities via TEO profile (bumps in within GCIs): (a) Normal Speech Waveform, (b) TEO profile for normal speech signal. Dysarthric Speech waveform for (c) severity-1, (e) severity-2, (g) severity-3, and (i) severity-4. TEO Profile of Dysarthric Speech for (d) severity-1, (f) severity-2, (h) severity-3, and (j) severity-4.

Therefore, the presence of bumps within two consecutive GCIs indicates the production of speech is not only due to the linear system, as in linear acoustics; rather, there is significant contribution from nonlinear effects such as aeroacoustic mechanisms (Quatieri, 2002; Teager & Teager, 1990). To that effect, Fig. 5 shows the TEO profile for the normal vs. dysarthic speech case with four severity levels. We can observe that bumps are present within two consecutive GCIs for both the normal and the dysarthric speech. Furthermore, as severity of dysarthria increases, Teager energy pulses corresponding to GCIs are irregularly located and, in addition, more high amplitude and noisy bumps can be noted, indicating much more significant nonlinear aspects in production of dysarthic speech.

3.3. Time–frequency analysis

Features derived from time–frequency representation of speech signal have been used in several speech applications. In particular, the study in Chen, Wang, and Wang (2014) evaluated various acoustic features based on their relative effectiveness to estimate quality of time–frequency mask, namely, ideal binary mask (IBM) — a central research issue in speech enhancement and source separation area. The wide range of acoustic features (primarily motivated by robust Automatic Speech Recognition (ASR)) such as Mel-Frequency Cepstral Coefficients (MFCC), Perceptual Linear Prediction, Relative Spectral Transform-Perceptual Linear Prediction, Gammatone Frequency Cepstral Coefficients, Power Normalized Cepstral Coefficients, fundamental frequency (F_0), etc. are considered in this study. In addition, study in Chen et al. (2014) proposed a new acoustic feature called the Multi-Resolution Cochleogram (MRCG), which is encoder multi-resolution power distribution in the time-frequency representation of a signal. Finally, study in Chen et al. (2014) found MRCG and pitch as complementary features using group Lasso (least absolute shrinkage selection operator) that improve l_1/l_2 mixed norm regularization on logistic regression to investigate the complementary features. The study in Chen et al. (2014) is extended in Delfarah and Wang (2017) for monaural speech separation from supervised learning perspective by predicting an ideal time–frequency mask from similar acoustic features of noisy speech under reverberant conditions at low signal-to-noise ratios (SNRs) and employing a simple Deep Neural Networks (DNNs) as a learning machine. The key findings of this study are that complementary feature sets for speech separation in reverberant conditions are different from those in anechoic conditions (as reported in Chen et al., 2014).

Motivated by these studies, we employ such representation for the dysarthic severity classification problem. Fig. 6-(a) and (b) show the plot of Short-Time Fourier Transform (STFT) vs. LP

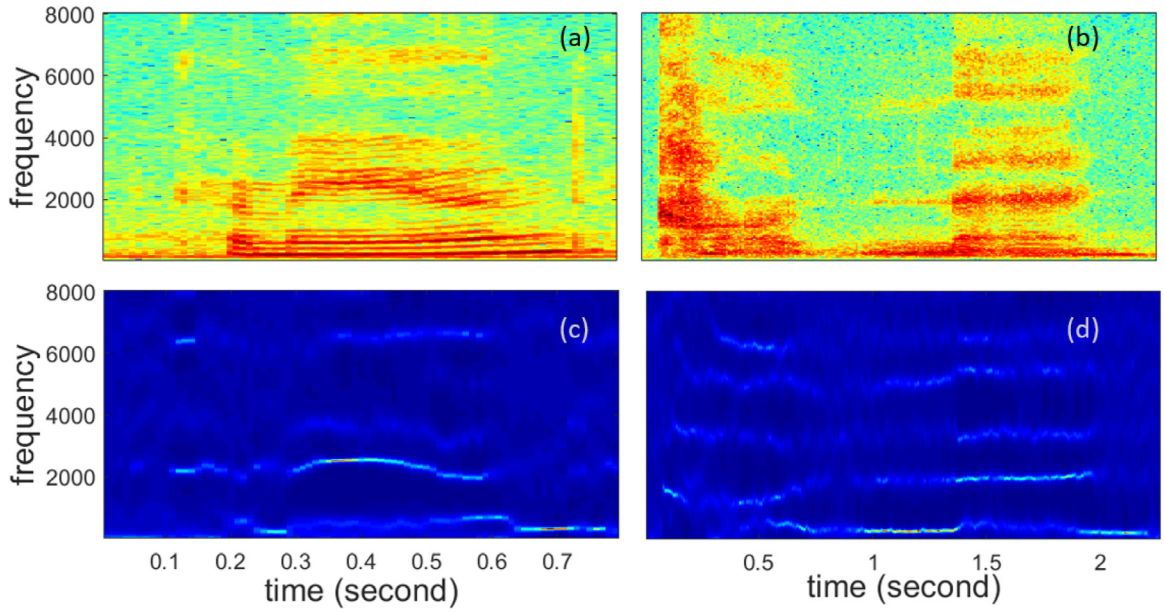


Fig. 6. STFT representation of: (a) Normal speech (b) Dysarthric speech *vs.* LP spectrum of (c) Normal speech (d) Dysarthric speech.

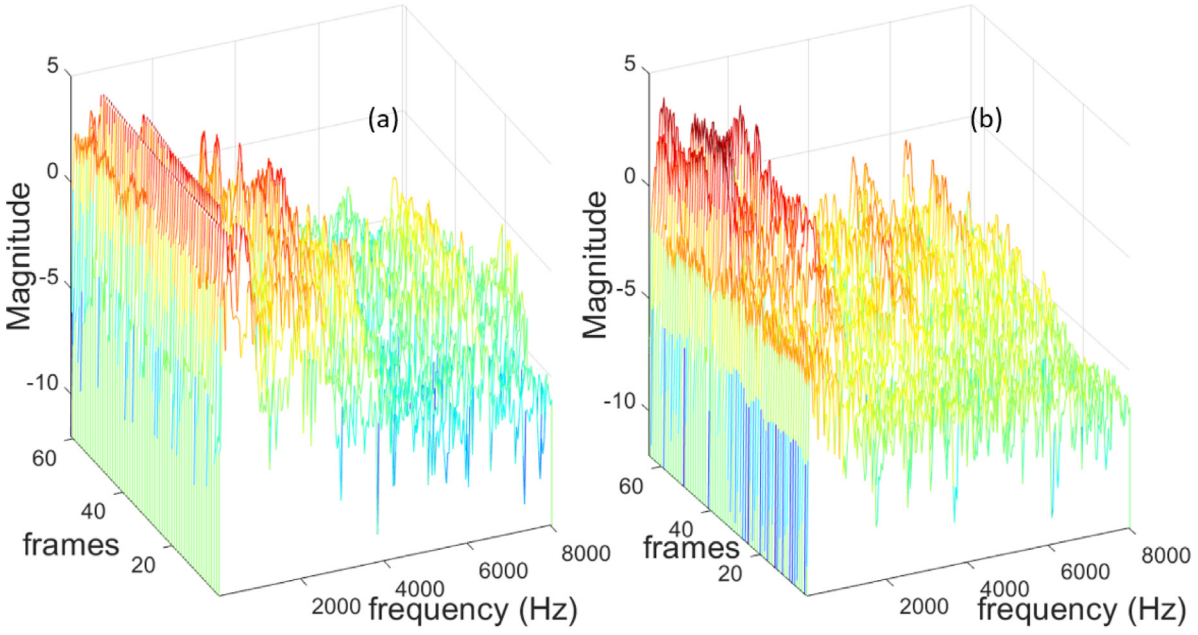


Fig. 7. Waterfall characteristics of: (a) Normal speech (b) Dysarthric speech.

spectrum for the normal vs. dysarthria speech case. We also show the waterfall plot in Fig. 7 to emphasize the corresponding joint time–frequency characteristics during the production of dysarthic speech. From the waterfall plots, we can observe that the formant structure is severely damaged for dysarthric speech as compared to its normal counterpart, where formant peaks and their evolving structures are clearly visible. Thus, the analysis presented in this section indicates that F_0, its harmonics, formants, and their structures are severely affected due to dysarthria, more so for high severity, and hence, we propose to exploit this unstructured spectral energy distribution captured via spectrograms as feature representation for the proposed deep learning architecture. In addition, TEO-based analysis helped us to observe relatively more severe nonlinearities in speech production. To that effect, authors believe that proposed deep learning-based architecture may help to imitate this nonlinearity effectively.

4. Proposed approach

In this section, we provide readers with a detailed description of the methodology and strategies used to solve the proposed problem. Specifically, as represented in Fig. 9, the following three major components exist:

1. onset–offset detection;
2. Time–Frequency (T–F) representation of selected short-duration speech segments;
3. mapping technique for utilizing features to do efficient classification.

Along with the schematic representation of the proposed methodology, we provide the respective Algorithm 1, used to solve our severity classification task. Detailed explanations can be found hereafter.

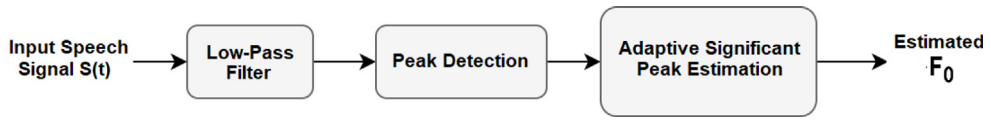


Fig. 8. Schematic representation of F_0 detection.
Source: Adapted from Bořil and Pollák (2004).

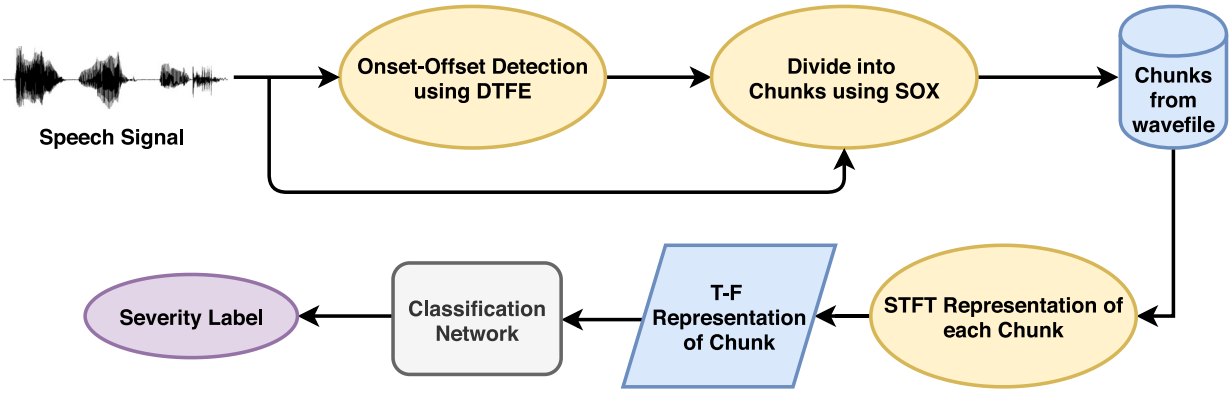


Fig. 9. Schematic representation of proposed methodology. After (Vásquez-Correa et al., 2017).

Algorithm 1 Proposed Algorithm for the Dysarthria Severity Classification using Short-duration Speech Segment

Result: Optimized weights and biases for Classifier and Severity label for the input Dysarthric Speech

for *number of iterations* **do**
  Randomly select a speech wave file s from the training set of the dysarthric speech $\mathcal{S}$;
  Detect the Onset-Offset from the wave file using DTFE method;
  Create a set of time stamps $\mathcal{T} = \{t_i\}_{i=1}^{n}$, where t_i represents the i^{th} seconds where Onset-Offset is detected;
  Remove all the $t \in \mathcal{T}$ where $t > 1second$;
  Create a set of chunks $\mathcal{C}$ with time-interval $(t - 100ms, t + 100ms)$, $\forall t \in \mathcal{T}$;
  Train the classification model:
  for $\forall c \in \mathcal{C}$ **do**
    spec = STFT(c, window-size=2ms, frame-shift=0.5ms);
    logits = $classification_{model}$(spec);
    $\hat{y}$ =argmax(logits);
    $\mathcal{L}(y, \hat{y})$ = CrossEntropyLoss(y, logits);
    Update the classifier by descending its stochastic gradient, i.e., $\nabla_{\theta_c}(\mathcal{L}(y, \hat{y}))$
  end
end

4.1. Onset–offset detection

The onset and offset regions of the speech signals were characterized, as a function of their fundamental frequencies (F_0), by using the Direct Time Fundamental Frequency Estimation (DTFE) method, described in a study reported in Bořil and Pollák (2004). DTFE is a novel algorithm for fundamental frequency (F_0) estimation performed directly in the time-domain. In this algorithm, F_0 detection is performed via evaluating actual F_0 candidate from the distance between neighboring significant peaks (i.e., local extremas) that there is only one peak representing the absolute maximum and one the absolute minimum in the quasi-period of the signal. Structure of the F_0 detection is shown in Fig. 8. Implementation details for pitch (F_0) tracker DTFE are presented here.[1]

We carefully used that method to extract the onset and offset time stamps from each input speech signal in our dataset. After this, the borders were detected and, in addition, 100 ms from each signal was taken to the left and 100 ms to the right of each border, forming the 200 ms-long signals as "chunks". Each one of those chunks was modeled by using the Short-Time Fourier Transform (STFT), as described ahead.

4.2. Spectrogram: T-F representation

As discussed in Section 3, STFT was applied to each generated chunk, for T-F representation. To feed the classifier, 2 ms-long frames, shifted 0.5 ms over time, were considered in order to generate a spectrogram image with dimensions of 570 × 450 pixels. Fig. 10 shows example spectrograms for different severity-levels of dysarthria. The spectrograms were plotted only for one second-long, i.e., short-duration, speech signals. Observably, the energy distribution across the frames, for speakers with different severity-levels of dysarthria, is significantly unlike. Hence, we hypothesize that those short-duration speech segments are sufficient for the intended classification. To support our hypothesis, we show the experimental results in Section 5.

4.3. Mapping technique: CNN vs. ResNet

A recent trend indicates that a high number of stacked layers in neural networks provides better results for classification task in general (LeCun, Bengio, & Hinton, 2015). Nevertheless, accuracy degrades rapidly after the increment in the number of layers. The reason behind this is the ample training error, instead of overfitting (He et al., 2016a). Moreover, current studies show that deep neural networks are more challenging to train due to overfitting, vanishing gradient, and besides additional issues, as explained in Angelov and Sperduti (2016), Goodfellow, Bengio, and Courville (2016) and Hestness, Ardalani, and Diamos (2019).

Making CNN models deeper for our task is not an appropriate solution. To overcome the limitation of CNN-based architectures,

[1] https://personal.utdallas.edu/~hynek/tools.html.

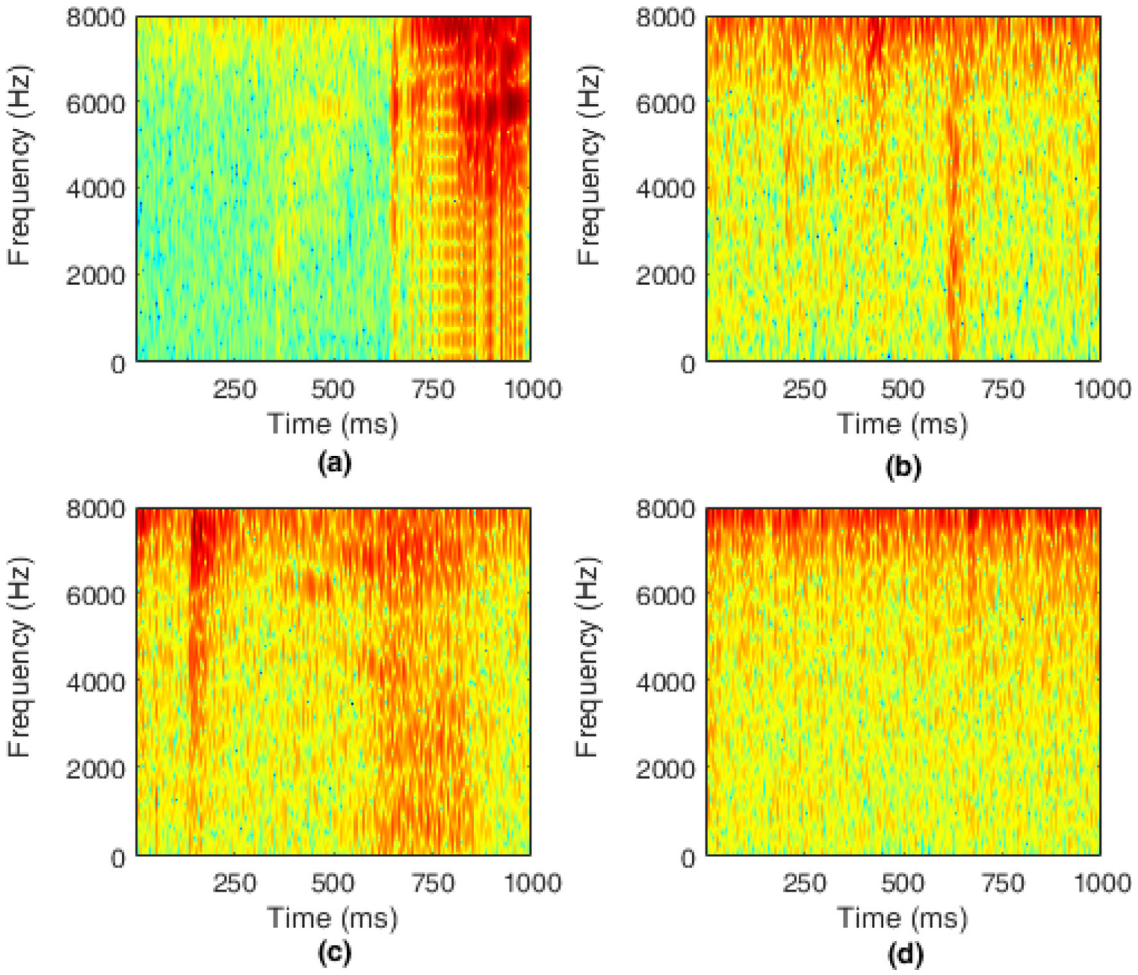


Fig. 10. STFT of the one second of speech segment of speakers with different severity-levels when they pronounce the word "Command": (a) Very Low, (b) Low, (c) Medium, and (d) High.

residual learning-based classifiers, ResNet, were used in He et al. (2016a) and Tai, Yang, and Liu (2017). The former technique is used as a baseline for comparisons. Although ResNet uses convolution layers as its building blocks, it is more effective than CNNs for image classification, as suggested in He et al. (2016a). Thus, we decided to use the strength of ResNet for our classification problem, analyzing its results. In residual-learning, $f(y)$ is the underlying function, as shown in Section 2, to be learned by a regular neural network-based classifier, where y is the set of input features, i.e., spectrogram image of chunk in our case. Due to the network non-linearity, it is capable of learning $f(y) - y$ along with $f(y)$, forcing the classifier to optimize the residual function $F(y) = f(y) - y$. Hence, the original optimization function becomes $F(y) + y$. Although both the methods are learning the same underlying functions, the ease of learning is different. The main reason behind that is the associated *identity mapping*, as shown in He, Zhang, Ren, and Sun (2016b). Due to the skip connections in ResNet, it would be easier to push the residual to zero than to fit an identity mapping by a stack of nonlinear layers. This helps ResNet in learning different patterns more efficiently, as shown in He et al. (2016a) and Tai et al. (2017).

5. Experimental setup and results

In this section, we provide the description about the database used for experiments, and the details of hyperparameters used both in the baseline system and in the proposed ResNet model. The results for the severity-based classification task for different architectures (GMM, CNN, LCNN, and ResNet) are discussed along with the analysis of the effectiveness of ResNet over other methods for our classification task.

5.1. Dataset

Universal Access (UA) corpus (Kim et al., 2008) was used in our experiments. This dataset includes details on speech intelligibility for each dysarthric speaker, in terms of severity-level, based on transcription tasks at the word-level performed by the human listeners. In our experiments, we used 8 speakers, i.e., 4 males namely M01, M05, M07, M09 and 4 females namely F02, F03, F04, F05. Details about them can be found in Kim et al. (2008). Each speaker produced a total of 765 isolated words, in which 455 words are distinct. For training and testing, we used 90% and 10% data, from 455 distinct words for each speaker, respectively.

5.2. Comparison methods

Our ResNet model has two types of residual blocks: (i) regular residual block; and (ii) downsampling-based residual block (He et al., 2016a). In this paper, our ResNet structure comprises nine regular and three downsampling-based residual blocks. In residual blocks, we used two 2-dimensional CNN layers with a kernel size of $3 \times 3 = 9$. However, for downsampling-based residual blocks, we increase the stride to 2 for the first CNN block and,

Table 2
Proposed architectural details of ResNet. Here, Conv1 and Conv2 show continuous layers of residual block and Conv3 shows parallel downsampling layer in residual block.

Block	# of Neurons	General settings	Conv1 settings	Conv2 settings	Conv3 settings
Convolution	3200	64, 7 × 7, 1	–	–	–
Batch normalization	–	64, –, –	–	–	–
Max pool	–	64, 7 × 7, 1	–	–	–
Residual block	1280	–	64, 3 × 3, 1	64, 3 × 3, 1	–
Residual block	1280	–	64, 3 × 3, 1	64, 3 × 3, 1	–
Residual block	1280	–	64, 3 × 3, 1	64, 3 × 3, 1	–
Residual down sampling	3840	–	128, 3 × 3, 2	128, 3 × 3, 1	128, 3 × 3, 2
Residual block	2560	–	128, 3 × 3, 1	128, 3 × 3, 1	–
Residual block	2560	–	128, 3 × 3, 1	128, 3 × 3, 1	–
Residual down sampling	7680	–	256, 3 × 3, 2	256, 3 × 3, 1	256, 3 × 3, 2
Residual block	5120	–	256, 3 × 3, 1	256, 3 × 3, 1	–
Residual block	5120	–	256, 3 × 3, 1	256, 3 × 3, 1	–
Residual down sampling	15 360	–	512, 3 × 3, 2	512, 3 × 3, 1	512, 3 × 3, 2
Residual block	10 240	–	512, 3 × 3, 1	512, 3 × 3, 1	–
Residual block	10 240	–	512, 3 × 3, 1	512, 3 × 3, 1	–
Average pool	–	512,8 × 8,–	–	–	–
Fully-connected layer	4	–	–	–	

before adding input x to the output of second CNN block, we use downsampling with similar settings as first block. Therefore, we use one downsampling-based shortcut connection with two residual blocks to process the downsampled output. Table 2 shows the architectural details of the proposed model. We first used a single CNN layer with 7 × 7 kernel size to downsample the input. Later, we used a total of 14 different residual blocks and, at the end, we adopted a single fully-connected layer with softmax activation function to predict the severity of the input dysarthric speech spectrogram. Architectural details related to the proposed ResNet are shown in Table 2.

For the baseline system, we have used Gaussian Mixture Model (GMM) as a classifier (Bishop, 2006; Duda & Hart, 2006; Reynolds, 1992; Reynolds & Rose, 1995). GMMs parameters are initialized with random initialization and updated based on Expectation Maximization (EM) algorithm. The parameters are updated up to 50 iterations. Four GMMs were trained for each severity level. Test sample was presented to each of the GMM to obtain its log-likelihood score (LLK). The GMM producing the maximum LLK is considered to be the predicted class.

We also designed a regular CNN-based architecture containing a 5 × 5 kernel for each one of its four CNN layers: CNN-layer-A, CNN-layer-B, CNN-layer-C and CNN-layer-D, with 8, 16, 32, and 64 output channels, respectively (LeCun, Bottou, Bengio, & Haffner, 1998). Moreover, we adopted max-pooling with a kernel size of 4 × 4 after the first three CNN blocks. Later, we used three fully-connected layers with 128, 64, and 4 output neurons. ReLU was used as an activation function for the hidden layers in both the models. Accordingly, the output layers in both the models are followed by a softmax activation function. The models were trained for 30 epochs with learning rate of 0.0001, by using Adam optimizer (Kingma & Ba, 2014).

The LCNN architecture was also employed here since it performed exceptionally when used for the spoof speech detection task (Lavrentyeva et al., 2017, 2019). Hence, we decided to use it for severity-based classification of dysarthric speech. These experiments were carried out by using spectrogram images of size 450 × 570 in Red-Green-Blue (RGB) color format. The LCNN architecture uses Max-Feature-Map (MFM) activation operation instead of other non-linearities, such as ReLU or sigmoid function. MFM activation is a special case of max-out function, for learning with a small number of parameters as compared with ReLU activation function (Wu, He, Sun, & Tan, 2018). Also, the MFM function has the better generalization ability for distinct data distributions. MFM function is defined as:

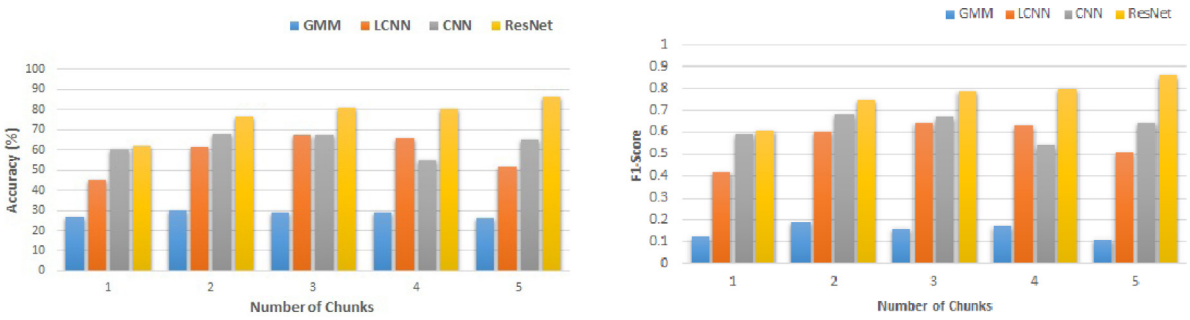
$$y_{ij}^{k} = max(x_{ij}^{k}, x_{ij}^{k+\frac{N}{2}}), \quad (1)$$

Table 3
Details of the proposed LCNN architecture for dysarthria severity classes.

Layer	Filter/Stride	Output	#Parameters
Conv1	5 × 5/1 × 1	32 × 450 × 570	2432
MFM1	–	16 × 450 × 570	–
MaxPool1	2 × 2/1 × 2	16 × 225 × 285	–
Conv2a	1 × 1/1 × 1	32 × 225 × 285	544
MFM2a	–	16 × 225 × 285	–
Conv2b	3 × 3/1 × 1	64 × 225 × 285	9280
MFM2b	–	32 × 225 × 285	–
MaxPool2	2 × 2/1 × 2	32 × 112 × 142	–
Conv3a	1 × 1/1 × 1	64 × 112 × 142	2112
MFM3a	–	32 × 112 × 142	–
Conv3b	3 × 3/1 × 1	128 × 112 × 142	36 992
MFM3b	–	64 × 112 × 142	–
MaxPool3	2 × 2/2 × 2	64 × 28 × 35	–
Conv4a	1 × 1/1 × 1	128 × 28 × 35	8320
MFM4a	–	64 × 28 × 35	–
Conv4b	3 × 3/1 × 1	64 × 28 × 35	36 928
MFM4b	–	32 × 28 × 35	–
MaxPool4	2 × 2/2 × 2	32 × 14 × 17	–
Conv5a	1 × 1/1 × 1	64 × 14 × 17	2112
MFM5a	–	32 × 14 × 17	–
Conv5b	3 × 3/1 × 1	32 × 14 × 17	9248
MFM5b	–	16 × 14 × 17	–
MaxPool5	2 × 2/2 × 2	16 × 7 × 9	–
FC6	–	1 × 128	24 704
MFM6	–	1 × 64	–
FC7	–	1 × 4	260

where the number of channels of the input convolution layer is $2N$, $(1 \leq k \leq N)$, $(1 \leq j \leq W)$, and $(1 \leq i \leq H)$. Here, i and j indicate the feature component and frame number, respectively. Each convolution layer is a combination of two independent terms previously calculated from input layer's output. The MFM activation function is used then to calculate element-wise maximum of those parts. Max-Pooling layers with kernel of size 2x2 and stride of size 2 × 2 were used for dimensionality reduction. The fully-connected FC6 layer contains a low-dimensional high-level audio representation. Then, the FC7 layer with softmax activation function was used to distinguish between four classes of dysarthric speech during the training process. The details of LCNN architecture are shown in Table 3.

As described in Section 4.1, 200 ms-long chunks were extracted from each speech signal. For our experiments, we selected a different number of onset-offset detection routines and then used them for training. In case the distance between consecutive onset-offset tags is less than 200 ms, we used overlapped chunks. In the case of non-overlapping chunks, we got [(number of chunks) × (200 ms)] seconds of speech segment, which



(a) accuracy (b) F1-score

Fig. 11. Baseline CNN *vs.* ResNet, for different speech-duration based on (a) classification accuracy score and (b) F1-Score. Additionally, LNCC and GMM were also considered for comparisons, however, since GMM exhibit a poor accuracy, its F1-scores were not even computed.

becomes, however, less than this in the case of overlapping scenarios. Hence, we took a maximum of [(number of chunks) × (200 ms)] seconds of speech from each utterance for training. The proposed system was assessed for different number of chunks, i.e., different speech-duration, to prove our hypothesis that maximum one-second of speech, i.e., five chunks, is a sufficient time for an efficient classification.

5.3. Performance evaluation

Accuracy and F1-score were used for performance evaluation, where the former is the number of correctly predicted wave files out of all the input wave files, and the latter is calculated by taking the harmonic mean of precision and recall for each class. Precision was considered as being the fraction of correct classified instances among all classifications for each class, and, in addition, recall was defined as being the fraction of correct classified instances among the ones that actually belong to that class. In particular, we calculated the F1-score for each class and presented the "macro average" results. We analyzed the performance of both systems in terms of different speech duration, i.e., maximum [(number of chunks) × (200 ms)] seconds. From Fig. 11-(a) and (b), we can clearly see that the proposed ResNet-based approach outperforms the baseline CNN. In particular, we got, on average, 21.35% and 22.48% of improvement compared with the baseline CNN in terms of classification accuracy and F1-score, respectively. For further comparisons, a Gaussian Mixture Model (GMM) and a Light-CNN (LCNN), which is a lighter version of the conventional CNN architecture, were also considered.

We can observe that GMM performed relatively poor compared to other systems, indicating its unsuitability for classifying severity of dysarthria from short-duration speech. Moreover, GMM is based on the first two moments only, i.e., mean and variance, which may not be adequate to represent nonlinearities in speech production mechanism, and more so for dysarthric speech, as discussed in Section 3. In addition, estimating higher-order moments with the same statistical confidence, as that of first two moments, requires a large amount of training data, which is not feasible in this problem due to the impossibility in getting long-duration dysarthric speech data. Contrary to this, deep learning architectures, in particular the proposed ResNet, are able to capture such nonlinearities from short-duration speech segments.

5.4. Analysis of results

In this subsection, we analyze how effective the proposed methodology is in two different aspects: (i) learning performance, and (ii) amount of training data. Since CNN and ResNet performed relative better than LCNN and GMM, as discussed in the previous subsection, we assess hereafter just the behavior of CNN and ResNet-based systems w.r.t. learning performance and the amount of training data.

5.4.1. Learning performance

To analyze the learning performance, we observed the output of the last layer just before the softmax activation from both of our architectures, i.e., CNN and ResNet. To analyze efficiently, we converted the image into binary format, where the white color part shows the pattern learned by the architecture, as illustrated in Fig. 12. For that analysis, we used *Guided Backpropagation Saliency* method in order to extract the region learned by any trained CNN-based classifier (Simonyan, Vedaldi, & Zisserman, 2013). In guided backpropagation, forward pass was performed till the target layer on input features is performed. Then, the disadvantageous neurons were kept to zero and back propagation was applied till the input features. More formally, the whole process can be explained as: (Springenberg, Dosovitskiy, Brox, & Riedmiller, 2014):

$$
\begin{aligned}
&\text{activation:} && f_i^{l+1} = relu(f_i^l) = max(f_i^l, 0)\\
&\text{backpropagation:} && R_i^l = (f_i^l > 0) \cdot R_i^{l+1}, \text{ where } R_i^{l+1} = \frac{\partial f^{out}}{\partial f_i^{l+1}}\\
&\text{backward 'deconvnet':} && R_i^l = (R_i^{l+1} > 0) \cdot R_i^{l+1}\\
&\text{guided backpropagation:} && R_i^l = (f_i^l > 0) \cdot (R_i^{l+1} > 0) \cdot R_i^{l+1} \quad .
\end{aligned}
$$

From Fig. 12, we can observe the advantage of ResNet over CNN for the dysarthric severity-based classification, and we can see that ResNet can learn various characteristics of dysarthric speech which are different from natural speech. To understand the advantage of ResNet in our problem, we explored the energy parameter. The energy in dysarthric speech is more distributed, i.e., energy fluctuations are more frequent, compared with natural speech (Kent, Weismer, Kent, Vorperian, & Duffy, 1999; Rudzicz, 2013). Moreover, it is observed that as the severity-level changes, the energy of dysarthric speech shows significant changes (Purohit, Parmar, Patel, Malaviya, & Patii, 0000). Hence, capturing these energy fluctuations from the spectrogram is an essential task. In other words, detecting patterns from low and high-frequency regions from the spectrogram can achieve this task and help the model to distinguish between different severity levels. Consequently, our short-duration speech segments include both the patterns (i.e., high and low energy regions), as shown in Fig. 10. From Panels I and ll, we can observe that ResNet is capturing both the regions efficiently, hence, performing better

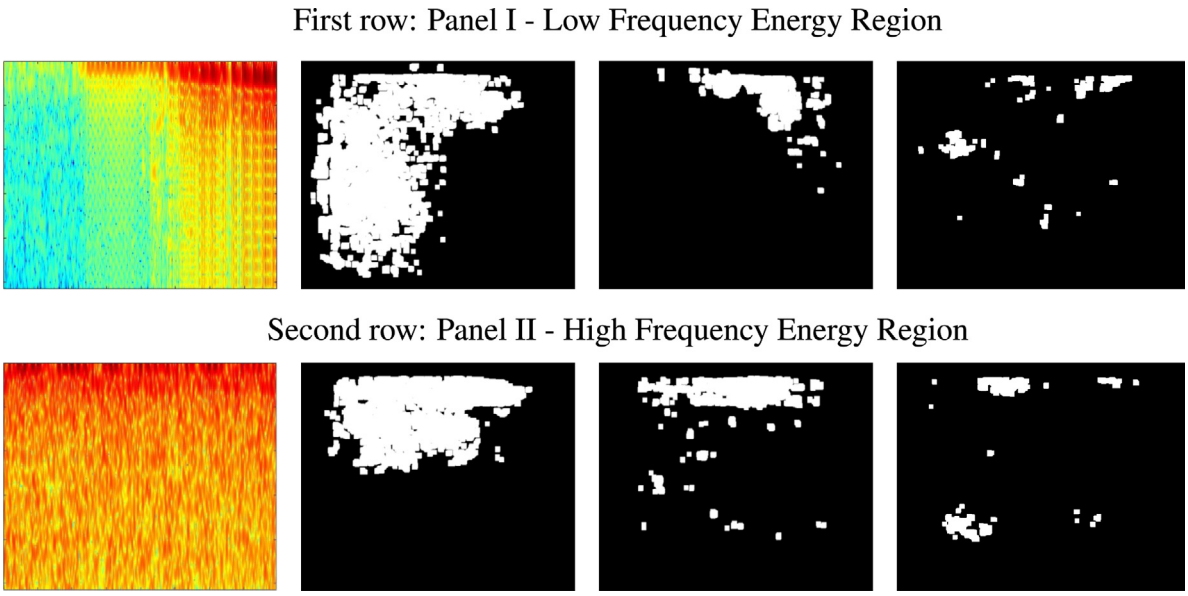
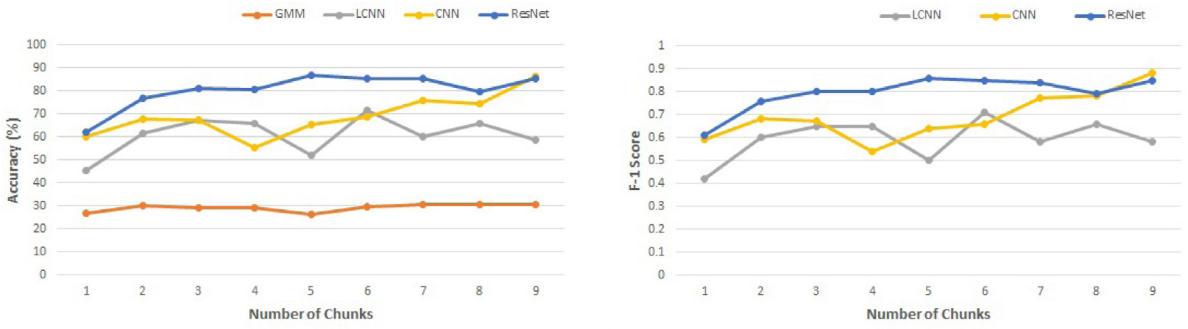


Fig. 12. Learning of proposed ResNet *vs.* baseline CNN. For both Panels, we have: [First Column]: Input spectrogram of chunk (horizontal axis: time, vertical axis: frequency); [Second Column]: Visualization of learning of ResNet; [Third Column]: Visualization of learning of CNN; [Fourth Column]: Visualization of learning of LCNN. Here, Visualization images are in the form of pixels.



(a) accuracy (b) F1-score

Fig. 13. Evaluation of baseline CNN *vs.* ResNet for different number of chunks (i.e., amount of training data) based on (a) classification accuracy score, and (b) F1-Score. As previously presented, GMM and LCNN were considered for comparisons, however, since GMM exhibit a poor accuracy, its F1-scores were not even computed.

compared to the CNN. In Fig. 12, CNN only captures high energy regions, however, fails to capture low energy regions, hence has poor performance compared to the ResNet.

5.4.2. Amount of training data

Here, we analyze the performance of both the systems w.r.t. the amount of training data. To do so, we increased the number of chunks one-by-one, i.e., increasing a speech-duration by maximum of 200 ms, as shown in Fig. 13. Observably, for five chunks, ResNet performance is high in terms of classification accuracy and F1-score. This analysis empirically supports our hypothesis that one second-long speech segments are sufficient for an efficient classification. With a maximum of one second-long speech segment, we got 86.63% classification accuracy and 0.86 F1-score for ResNet. Contrary to this, we obtained 64.35% classification accuracy and 0.64 F1-score for CNN, respectively.

In complement, as the goal of this work is to detect dysarthria severity-level by using short-segments of speech, we also analyzed both ResNet and CNN structures when the entire speech utterance is available for training and testing. As shown in Table 4, ResNet exhibits superior performance for this task. Hence,

Table 4
Evaluation of baseline CNN *vs.* ResNet when entire speech utterance is available for training.

Systems	Accuracy (%)	F1-Score
ResNet	**98.90**	**0.98**
CNN	91.76	0.91

ResNet outperforms baseline CNN for both the classification scenarios, i.e., using short-duration speech segments and using entire speech utterances. This definitively proves ResNet is superior for the intended classification task.

5.5. Complementary comments

Additionally, we can observe that LCNN could capture both high and low frequency regions, however, the capture ratio is significantly lower compared to ResNet and to CNN, as also shown in Fig. 12. In contrast, ResNet is efficient in capturing both regions. Therefore, our ResNet structure not only outperforms the baseline CNN, but also LCNN and GMM. Observably, GMM was the worst classifier in terms of accuracy.

In addition to the mentioned comparison methods, we have also investigated other variants of the ResNet architecture, in

particular, ResNeSt, which uses a Split Attention layer between the two convolutional layers in the ResNet block. However, results obtained were poorer than the original ResNeSt architecture. Therefore, it has not been included in the manuscript.

6. Summary and conclusions

In this paper, a novel technique to detect dysarthria severity-levels was proposed. In particular, we presented time-domain, frequency-domain, and TEO analysis of dysarthric speech to justify spectrogram as feature representation particularly capable of capturing unstructured spectral energy density distributions. Additionally, we investigated nonlinearities in production of dysarthric speech using TEO. Our results indicate that GMM performs poorly than other systems, suggesting deep learning-based architectures and, in particular, the proposed ResNet. Based on short-duration speech segments and ResNets, our strategy differs from current state-of-the-art methods, in which long-duration speech segments feed a CNN. Our relevant experiments show that the former classifier outperforms the latter in terms of accuracy and F1-score, not only for short-speech segments but also for long ones. We observed, however, that only ResNet succeed in using short speech tags to detect dysarthria severity-levels.

Although our method shows remarkable results for the intended goal, the detection of onset-offset points and the subsequent spectrogram characterizations, both in real-time, are time-consuming. Hence, real-time implementation of this system in limited-capacity devices is still a challenge. Nevertheless, since our initial hypotheses were confirmed, we are satisfied with the results. Notably, the proposed approach opens a large and promising source of possibilities to explore the application of ResNets in speech processing and biomedical sciences.

In the future, we will try additional feature extraction techniques, such as Mel Cepstral Coefficients (MCCs), and slightly modified versions of ResNet to allow for more modest hardware requirements, consequently making the system more adequate for real-time implementations.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

R.C. Guido gratefully acknowledges the grants provided by the Brazilian agencies “National Council for Scientific and Technological Development (CNPq)” and “The State of São Paulo Research Foundation (FAPESP)”, Brazil, respectively through the processes 306808/2018-8 and 2019/04475-0, in support of this research.

References

An, K., Kim, M. J., Teplansky, K., Green, J. R., Campbell, T., Yunusova, Y., et al. (2018). Automatic early detection of amyotrophic lateral sclerosis from intelligible speech using convolutional neural networks. In *Proc. Interspeech* Hyderabad, India: (pp. 1913–1917).

Ananthapadmanabha, T. V., & Yegnanarayana, B. (1975). Epoch extraction of voiced speech. *IEEE Transactions on Acoustics, Speech, and Signal Processing, ASSP-23*(6), 562–570.

Ananthapadmanabha, T. V., & Yegnanarayana, B. (1979). Epoch extraction from linear prediction residual for identification of closed glottis interval. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 27*(4), 309–319.

Angelov, P., & Sperduti, A. (2016). Challenges in deep learning. In *The 24th European Symposium on Artificial Neural Networks (ESANN)* Bruges, Belgium: (pp. 489–496).

Atal, B. S., & Hanauer, S. L. (1971). Speech analysis and synthesis by linear prediction of the speech wave. *Journal of the Acoustical Society of America, 50*(2), 637–655.

Bhat, C., Das, B., Vachhani, B., & Kopparapu, S. K. (2018). Dysarthric speech recognition using time-delay neural network based denoising autoencoder. In *INTERSPEECH* Hyderabad, India: (pp. 451–455).

Bhat, C., Vachhani, B., & Kopparapu, S. K. (2016). Recognition of dysarthric speech using voice parameters for speaker adaptation and multi-taper spectral estimation. In *INTERSPEECH* San Fransisco, USA: (pp. 228–232).

Bhat, C., Vachhani, B., & Kopparapu, S. K. (2017a). Automatic assessment of dysarthria severity level using audio descriptors. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, USA* (pp. 5070–5074).

Bhat, C., Vachhani, B., & Kopparapu, S. K. (2017b). Automatic assessment of dysarthria severity level using audio descriptors. In *Proc. ICASSP* New Orleans, USA: (pp. 5070–5074).

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Bořil, H., & Pollák, P. (2004). Direct time domain fundamental frequency estimation of speech in noisy conditions. In *12th European Signal Processing Conference (EUSIPCO)* Vienna, Austria: (pp. 1003–1006).

Calvert, G., Spence, C., Stein, B. E., et al. (2004). *The Handbook of Multisensory Processes*. MIT Press, Edition.

Chandrashekar, H. M., Karjigi, V., & Sreedevi, N. (2019). Spectro-temporal representation of speech for intelligibility assessment of dysarthria. *IEEE Journal of Selected Topics in Signal Processing, 14*(2), 390–399.

Chen, J., Wang, Y., & Wang, D. (2014). A feature study for classification-based speech separation at low signal-to-noise ratios. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 22*(12), 1993–2002.

Chen, Z., Xie, Z., Zhang, W., & Xu, X. (2017). Resnet and model fusion for automatic spoofing detection. In *INTERSPEECH* (pp. 102–106).

kyong Choe, Y., Liss, J. M., Azuma, T., & Mathy, P. (2012). Evidence of cue use and performance differences in deciphering dysarthric speech. *Journal of the Acoustical Society of America, 131*(2), EL112–EL118.

Christensen, O., & Christensen, K. L. (2006). *Approximation Theory: From Taylor Polynomials to Wavelets*. Birkhauser.

Connaghan, K. P., Wertheim, C., Laures-Gore, J. S., Russell, S., & Patel, R. (2020). An exploratory study of student, speech-language pathologist and emergency worker impressions of speakers with dysarthria. *International Journal of Speech-Language Pathology, 14*, 1–10.

De Russis, L., & Corno, F. (2019). On the impact of dysarthric speech on contemporary ASR cloud platforms. *Journal of Reliable Intelligent Environments, 5*(3), 163–172.

Delfarah, M., & Wang, D. (2017). Features for masking-based monaural speech separation in reverberant conditions. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25*(5), 1085–1094.

Duda, R. O., & Hart, P. E. (2006). *Pattern classification*. John Wiley & Sons.

Enderby, P. (1980). Frenchay dysarthria assessment. *British Journal of Disorders of Communication, 15*(3), 165–173.

Fahn, S., & Elton, R. (2000). Unified Parkinson’s disease rating scale (UPDRS). *Revue Neurologique (Paris), 156*, 534–541.

Falk, T. H., Chan, W.-Y., & Shein, F. (2012). Characterization of atypical vocal source excitation, temporal dynamics and prosody for objective measurement of dysarthric word intelligibility. *Speech Communication, 54*(5), 622–631.

Farhadipour, A., Veisi, H., Asgari, M., & Keyvanrad, M. A. (2018). Dysarthric speaker identification with different degrees of dysarthria severity using deep belief networks. *ETRI Journal, 40*(5), 643–652.

Freed, D. B. (2018). *Motor Speech Disorders: Diagnosis and Treatment* (3rd ed.). Plural Publishing, USA.

Giri, M. P., & Rayavarapu, N. (2018). Assessment on impact of various types of dysarthria on acoustic parameters of speech. *Journal of the Acoustical Society of America, 21*(3), 705–714.

Gomez, P., et al. (2019). Characterization of Parkinson’s disease dysarthria in terms of speech articulation kinematics. *Biomedical Signal Processing and Control, 52*, 312–320.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Gurevich, N., & Scamihorn, S. L. (2017). Speech-language pathologists’ use of intelligibility measures in adults with dysarthria. *American Journal of Speech-Language Pathology, 26*(3), 873–892.

He, K., Zhang, X., Ren, S., & Sun, J. (2016a). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, Nevada: (pp. 770–778).

He, K., Zhang, X., Ren, S., & Sun, J. (2016b). Identity mappings in deep residual networks. In *European Conference on Computer Vision (ECCV)* Amsterdam, The Netherlands: (pp. 630–645).

Hestness, J., Ardalani, N., & Diamos, G. (2019). Beyond human-level accuracy: computational challenges in deep learning. In *The 24th Symposium on Principles and Practice of Parallel Programming* (pp. 1–14).

Hoehn, M. M., & Yahr, M. D. (1967). Parkinsonism: Onset, progression, and mortality. *Neurology, 17*(5), 427.

Huang, G., Liu, Z., Weinberger, K. Q., & van der Maaten, L. (2017). Densely connected convolutional networks. In *Proc. Computer Vision and Pattern Recognition (CVPR) Cnonference* Honolulu, HI, USA: (pp. 2261–2269).

Jiang, Y., Chen, L., Zhang, H., & Xiao, X. (2019). Breast cancer histopathological image classification using convolutional neural networks with small SE-ResNet module. *PloS One*, *14*(3), Article e0214587.

Jung, H., Choi, M.-K., Jung, J., Lee, J.-H., Kwon, S., & Young Jung, W. (2017). ResNet-based vehicle classification and localization in traffic surveillance systems. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 61–67).

Kaiser, J. F. (1990). On a simple algorithm to calculate the energy of a signal. In *Proc. of Int. Conf. on Acoustic, Speech and Signal Processing* (pp. 381–384).

Kawaguchi, K., & Bengio, Y. (2019). Depth with nonlinearity creates no bad local minima in resnets. *Neural Networks*, *118*, 167–174.

Kent, R. D., Weismer, G., Kent, J. F., Vorperian, H. K., & Duffy, J. R. (1999). Acoustic studies of dysarthric speech: Methods, progress, and potential. *Journal of Communication Disorders*, *32*(3), 141–186.

Kim, M. J., Cao, B., An, K., & Wang, J. (2018). Dysarthric speech recognition using Convolutional LSTM neural network. In *INTERSPEECH* Hyderabad, India: (pp. 2948–2952).

Kim, H., Hasegawa-Johnson, M., & Perlman, A. (2011). Vowel contrast and speech intelligibility in dysarthria. *Folia Phoniatrica et Logopaedica*, *63*(4), 187–194.

Kim, H., Hasegawa-Johnson, M., Perlman, A., Gunderson, J., Huang, T. S., Watkin, K., et al. (2008). Dysarthric speech database for universal access research. In *INTERSPEECH* Brisbane, Australia: (pp. 1741–1744).

Kim, Y. J., Kent, R. D., & Weismer, G. (2011). An acoustic study of the relationships among neurologic disease, dysarthria type and severity of dysarthria. *Journal of Speech, Language, and Hearing Research*, *54*(2), 417–429.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, {Last Accessed: Jan 30, 2017}.

Lansford, K. L., Berisha, V., & Utianski, R. L. (2016). Modeling listener perception of speaker similarity in dysarthria. *Journal of the Acoustical Society of America*, *139*(6), EL209–EL215.

Lansford, K. L., & Liss, J. M. (2014). Vowel acoustics in dysarthria: Speech disorder diagnosis and classification. *Journal of Speech, Language, and Hearing Research*, *57*(1), 57–67.

Lavrentyeva, G., Novoselov, S., Malykh, E., Kozlov, A., Kudashev, O., & Shchemelinin, V. (2017). Audio replay attack detection with deep learning frameworks. In *INTERSPEECH* Stockholm, Sweden: (pp. 82–86).

Lavrentyeva, G., Novoselov, S., Tseren, A., Volkova, M., Gorlanov, A., & Kozlov, A. (2019). STC antispoofing systems for the ASVspoof2019 challenge. In *INTERSPEECH* Graz, Austria: (pp. 1033–1037).

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553).

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Liu, S., Tian, G., & Xu, Y. (2019). A novel scene classification model combining resnet based transfer learning and data augmentation with a filter. *Neurocomputing*, *338*, 191–206.

Liu, X., Zhou, Y., Zhao, J., Yao, R., Liu, B., Ma, D., et al. (2020). Multiobjective resnet pruning by means of EMOAs for remote sensing scene classification. *Neurocomputing*, *381*, 298–305.

Lu, Z., Jiang, X., & Kot, A. (2018). Deep coupled resnet for low-resolution face recognition. *IEEE Signal Processing Letters*, *25*(4), 526–530.

Mustafa, M. B., Salim, S. S., Mohamed, N., Al-Qatab, B., & Siong, C. E. (2014). Severity-based adaptation with limited data for ASR to aid dysarthric speakers. *Public Library of Science (PLoS) One*, *9*(1).

de Oliveira Chappaz, R., dos Santos Barreto, S., & Ortiz, K. Z. (2018). Pneumo-phono-articulatory coordination assessment in dysarthria cases: a cross-sectional study. *São Paulo Medical Journal*, *136*(3), 216–221.

Paja, M. S., & Falk, T. H. (2012a). Automated dysarthria severity classification for improved objective intelligibility assessment of spastic dysarthric speech. In *Proc. INTERSPEECH* Portland, Oregon, (pp. 62–65).

Paja, M. O. S., & Falk, T. H. (2012b). Automated Dysarthria Severity Classification for Improved Objective Intelligibility Assessment of Spastic Dysarthric Speech. In *Proc. Interspeech* Portland, OR, USA: (pp. 62–65).

Perez, M., Aldeneh, Z., & Provost, E. M. (2020). Aphasic speech recognition using a mixture of speech intelligibility experts. In *Proc. Interspeech 2020* (pp. 4986–4990). http://dx.doi.org/10.21437/Interspeech.2020-2049.

Purohit, M., Parmar, M., Patel, M., Malaviya, H., & Patii, H. A. (0000). Weak speech supervision: A case study of dysarthria severity classification. In *2020 28th European Signal Processing Conference (EUSIPCO)* IEEE; (pp. 101–105).

Quatieri, T. F. (2002). *Discrete-Time Speech Signal Processing: Principles and Practices*. Pearson Education.

Reynolds, D. A. (1992). *A Gaussian mixture modeling approach to text-independent speaker identification* (Ph.D. thesis), Georgia Institute of Technology.

Reynolds, D. A., & Rose, R. C. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, *3*(1), 72–83.

Rosen, K. M., Goozee, J. V., & Murdoch, B. E. (2008). Examining the effects of multiple sclerosis on speech production: Does phonetic structure matter? *Journal of Communication Disorders*, *41*(1), 49–69.

Rudzicz, F. (2013). Adjusting dysarthric speech signals to be more intelligible. *Computer Speech & Language*, *27*(6), 1163–1177.

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.

Schmitz-Hübsch, T., Du Montcel, S. T., Baliko, L., Berciano, J., Boesch, S., Depondt, C., et al. (2006). Scale for the assessment and rating of aataxia: Development of a new clinical scale. *Neurology*, *66*(11), 1717–1720.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034, {Last Accessed: Apr 19, 2014}.

Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. (2014). Striving for simplicity: The all convolutional net. arXiv preprint arXiv:1412.6806.

Swarup, P., Maas, R., Garimella, S., Mallidi, S. H., & Hoffmeister, B. (2019). Improving ASR confidence scores for alexa using acoustic and hypothesis embeddings. In *INTERSPEECH*, Graz, Austria: (pp. 2175–2179).

Sztahó, D., & Vicsi, K. (2016). Estimating the severity of Parkinson's disease using voiced ratio and nonlinear parameters. In P. Král, & C. Martín-Vide (Eds.), *Statistical Language and Speech Processing* (pp. 96–107). Springer International Publishing.

Tai, Y., Yang, J., & Liu, X. (2017). Image super-resolution via deep recursive residual network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* Honolulu, Hawaii: (pp. 3147–3155).

Teager, H. M., & Teager, S. M. (1990). Evidence for nonlinear sound production mechanisms in the vocal tract. In *Chapter "Evidence for nonlinear sound production mechanisms in the vocal tract"* (pp. 241–261). Kluwer Academic Publishers.

Tripathi, A., Bhosale, S., & Kopparapu, S. K. (2020). Improved speaker independent dysarthria intelligibility classification using deepspeech posteriors. In *Proc. ICASSP* Barcelona, Spain: (pp. 6114–6118).

Turner, G., Tjaden, K., & Weismer, G. (1995). The influence of speaking rate on vowel space and speech intelligibility for individuals with amyotrophic lateral sclerosis. *Journal of Speech, Language, and Hearing Research*, *38*(5), 1001–1013.

Vachhani, B., Bhat, C., & Kopparapu, S. K. (2018). Data augmentation using healthy speech for dysarthric speech recognition. In *Proc. INTERSPEECH* Hyderabad, India: (pp. 471–475).

Vásquez Correa, J. C., Arias, T., Orozco-Arroyave, J. R., & Nöth, E. (2018). A multitask learning approach to assess the dysarthria severity in patients with Parkinson's disease. In *Proc. Interspeech* (pp. 456–460).

Vásquez-Correa, J. C., Arias-Vergara, T., Orozco-Arroyave, J. R., & Nöth, E. (2018). A multitask learning approach to assess the dysarthria severity in patients with Parkinson's Disease. In *INTERSPEECH*, Hyderabad, India: (pp. 456–460).

Vásquez-Correa, J. C., Orozco-Arroyave, J. R., & Nöth, E. (2017). Convolutional neural network to model articulation impairments in patients with Parkinson's Disease. In *INTERSPEECH* Stockholm, Sweden: (pp. 314–318).

Vydana, H. K., & Vuppala, A. K. (2017). Residual neural networks for speech recognition. In *2017 25th European Signal Processing Conference (EUSIPCO)* (pp. 543–547). IEEE.

Wang, D., & Chen, J. (2018). Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(10), 1702–1726.

Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., et al. (2017). Residual attention network for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3156–3164).

Watanabe, S., Arasaki, K., Nagata, H., & Shouji, S. (1994). Analysis of dysarthria in amyotrophic lateral sclerosis—MRI of the tongue and formant analysis of vowels. *Rinsho Shinkeigaku*, *34*(3), 217–223.

Wu, X., He, R., Sun, Z., & Tan, T. (2018). A light CNN for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, *13*(11), 2884–2896.

Yang, S., Wang, F., Yang, L., Xu, F., Luo, M., Chen, X., et al. (2020). The physical significance of acoustic parameters and its clinical significance of dysarthria in Parkinson's disease. *Scientific Reports*, *10*(1), 11776.

Yorkston, K. M., Beukelman, D. R., & Traynor, C. (1984a). *Assessment of Intelligibility of Dysarthric Speech*. Pro-ed Austin, TX.

Yorkston, K., Beukelman, D., & Traynor, C. (1984b). *Computerized Assessment of Intelligibility of Dysarthic Speech*. CC Publications.

Young, V., & Mihailidis, A. (2010). Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review. *Assistive Technology*, *22*(2), 99–112.

Zhang, X., & Wang, D. (2016). Boosting contextual information for deep neural network based voice activity detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *24*(2), 252–264.

Zhang, X., & Wu, J. (2013). Deep belief networks based voice activity detection. *IEEE Transactions on Audio, Speech, and Language Processing*, *21*(4), 697–710.