Speech Communication 147 (2023) 1–11

Contents lists available at ScienceDirect

Speech Communication

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

Dysarthria severity classification using multi-head attention and multi-task learning

Amlu Anna Joshy *, Rajeev Rajan[1]

College of Engineering Trivandrum, APJ Abdul Kalam Technological University, Thiruvananthapuram 695016, Kerala, India

ARTICLE INFO

Keywords:
Dysarthria
Multi-head attention
Multi-task learning
Convolutional neural network

ABSTRACT

Identifying the severity of dysarthria is considered a diagnostic step in monitoring the patient's progress and a beneficial step in the transcription of dysarthric speech. In this paper, the effectiveness of using the multi-head attention mechanism (MHA) and the multi-task learning approach is explored for automated dysarthria severity level classification. Dysarthric speech utterances are represented by mel spectrograms and fed to a residual convolutional neural network for effective feature learning. Then the MHA module is added to identify the salient severity-highlighting periods. At the classification end, gender, age, and disorder-type identifications are employed as auxiliary tasks to share mutual information and leverage the severity classification. The performance of the proposed method is evaluated on the Universal Access Speech database. By giving a gain of 11.51% classification accuracy over the baseline system under the speaker-dependent scenario and 11.58% under the speaker-independent scenario, the proposed system demonstrates its potential for the dysarthria severity classification.

1. Introduction

Dysarthria is a motor speech disorder that is either acquired due to a neurological injury such as cerebral palsy (CP), or developed as a symptom of any neuro-degenerative diseases such as Parkinson's disease (PD) and amyotrophic lateral sclerosis (ALS) (Rudzicz, 2010). The motor speech sub-systems get impaired or are weakly coordinated, resulting in improper speech production. This causes slurred speech, variable speaking rates, and irregular phoneme articulations, which in turn deteriorate the speech quality. It can greatly hinder people from effectively communicating with others, as the speech intelligibility would be reduced partially or completely (Kent et al., 1999). Also, other perceptual attributes of speech vary with the dysarthria severity. Severe dysarthrics would have shorter tone units and higher mean fundamental frequencies than mild dysarthrics, whose speech would be more 'monotonous' (Schlenck et al., 1993). Progressive dysarthria as seen in PD patients can lead to progressive decline in muscle functioning over time (Qualls and Battle, 2012). Therefore it is important to monitor the progression of dysarthria, based on which the medication and speech therapy sessions are chosen. There are standard methods to assess dysarthria severity level such as, the dysarthria profile (Robertson, 1982), the Frenchay dysarthria assessment (FDA) (Enderby, 1980) and the dysarthria examination battery (DEB) (Drummond, 1993). Intelligibility rate is one of the factors in these methods, and it evidently shows the severity level.

A perceptual evaluation of speech intelligibility is usually done by a trained speech-language pathologist (SLP) using methods such as the percentage of consonants correct, which is defined as the ratio of the number of correctly uttered consonants to the number of total consonants in words or sentences (Shriberg and Kwiatkowski, 1982). This assessment would be inconsistent due to the familiarity of the SLP with the patient and would vary across clinicians with experience and listening skills. This demands the need for an automatic dysarthria severity level classification system. Such a system would be economical, consistent, and can be used for remote patient rehabilitation. Dysarthric patients have physical incapacities such as trembling hands, due to the weak coordination of muscles, which make the use of a keyboard or a joystick-based interactive application less useful for their communication purposes. Speech recognisers specifically designed for them can potentially be benefited from an effective automatic dysarthria severity level classification approach.

* Correspondence to: Electronics and Communication Engineering Department, College of Engineering Trivandrum, APJ Abdul Kalam Technological University, Thiruvananthapuram, India.

E-mail addresses: amluanna02@gmail.com (A.A. Joshy), rajeev@cet.ac.in (R. Rajan).

[1] Rajeev Rajan was with the Speech and Music Technology Lab, Department of Computer Science and Engineering, Indian Institute of Technology, Madras, India. He is now with the Electronics and Communication Engineering Department, College of Engineering Trivandrum, APJ Abdul Kalam Technological University, Thiruvananthapuram, India.

https://doi.org/10.1016/j.specom.2022.12.004
Received 5 May 2022; Received in revised form 12 November 2022; Accepted 27 December 2022
Available online 30 December 2022

1.1. Objective intelligibility assessment of dysarthria

For implementing a reliable objective assessment method, proper acoustic features have to be selected to capture the discriminative intelligibility characteristics. The type and severity of dysarthria can also be identified from acoustic features. Intelligibility rate is one of the factors in these methods, and it evidently shows the severity level. In the literature, pathological speech intelligibility assessment in general, and detection and severity identification of dysarthria in specific are done using different approaches.

In Kadi et al. (2013) prosodic features like mean pitch, jitter, shimmer, articulation rate, the proportion of the vocalic duration, harmonics to noise ratio, and degree of voiced breaks are selected by linear discriminant analysis (LDA) and classification of dysarthria severity levels is performed by two approaches, namely a Gaussian mixture model (GMM) and a support vector machine (SVM). The same authors have proposed computational models to represent the auditory perceptual knowledge in Kadi et al. (2016) by simulating the external, middle and inner parts of the ear. The obtained auditory-based cues were combined with the mel frequency cepstral coefficients (MFCC) and fed to GMM, SVM, hybrid GMM/SVM classifiers for dysarthric speaker identification and severity assessment. Another feature selection technique using a genetic algorithm, from speech disorder-specific prosodic features like spectral moments, formants, skewness and MFCCs is proposed in Vyas et al. (2016). Detection and severity classification of dysarthria is done here by an SVM classifier. Other complex features are also explored in literature such as, the breathiness indices in Chandrashekar et al. (2019a) and glottal parameters along with the openSMILE-based acoustic features in Prabhakera and Alku (2018). These works also used an SVM classifier. Audio descriptors extracted using multi-tapered spectral estimation technique are employed in Bhat et al. (2017) with an artificial neural network (ANN) at the classifier side. Perceptual linear prediction (PLP) features along with the energy component and moments are used in Martínez et al. (2015) to investigate the i-vector subspace modelling for intelligibility assessment of dysarthric speech. A linear predictor and a support vector regression predictor are compared at the predictor side. Perceptually enhanced single frequency filtering based cepstral coefficients (PE-SFCC) are proposed in Gurugubelli and Vuppala (2019) for severity classification. This employed the concept of single frequency filtering (SFF) for feature extraction, and the i-vector subspace modelling with the probabilistic linear discriminant analysis (PLDA) for classification.

While the above-mentioned works used machine learning classifiers, more recent works concentrate on improving the performance of deep learning models at the classification side too. MFCCs and log filter banks are compared against i-vectors in Bhat and Strik (2020) with a bidirectional long short-term memory (BLSTM) network classifier for identifying the dysarthric utterances from the healthy ones. Transfer learning was explored in this regard to improve the system performance. An attention-based LSTM model has been proposed in Millet and Zeghidour (2019) for the same task. Here the neural network jointly learns a filterbank, a normalisation factor and a compression power from the raw speech waveforms, together with the model architecture. Severity assessment of dysarthria has been done using basic acoustic features such as MFCCs along with their statistical moments in Joshy and Rajan (2021). Deep neural network (DNN), convolutional neural network (CNN) and LSTM-based recurrent neural networks (RNNs) are analysed at the classifier side against SVM. The impact of speech-disorder specific features and i-vectors are studied for the same task in Joshy and Rajan (2022). These works prove the efficacy of the deep learning models against the machine learning classifier. Two different novel strategies have been adopted in Tripathi et al. (2020a,b), to implement speaker independent (SID) intelligibility classification and dysarthria severity assessment systems using the features obtained from DeepSpeech, an end-to-end deep learning-based speech-to-text engine.

When the above-mentioned works used acoustic features at the input side, there are works using time–frequency representations as well. Log mel spectrograms have been used for the classification of speech into ALS, PD or healthy, and further for the specific severity classification of ALS and PD in Suhas et al. (2020). They have highlighted the fact that spectrograms work better than MFCCs on CNN models for the proposed task. Mel scale spectrograms along with their deltas are used with time-CNN and frequency-CNN models to capture temporal and spectral variations separately for the early detection of ALS in An et al. (2018). In contrast to this, joint spectro-temporal features from mel scale spectrogram are used in Chandrashekar et al. (2019b) for dysarthric speech intelligibility assessment. Two-dimensional discrete cosine transform (2D-DCT) coefficients extracted from mel scale spectrogram were used with ANN, and spectrograms in different forms were fed to CNN classifiers. This study analysed the performance of a time–frequency CNN architecture against a time-CNN and a frequency CNN. Their results revealed that the joint modelling of spectral and temporal information by the former model works better than the latter models which capture only one among these.

The authors have done another detailed investigation on the different time–frequency representations (TFRs) in Chandrashekar et al. (2020). TFRs like short-time Fourier transform (STFT) and SFF, with and without mel scale warping followed by perceptual enhancement were studied. The resulting spectrograms were compared with constant-Q transform (CQT) spectrograms and it was reported that the latter is better among all the spectrograms studied, at the cost of computational time. These works use light CNN architectures, and they have proved to be better classifiers than ANNs due to their representation learning capabilities. Spectrograms of short speech segments have been used with residual neural networks (ResNet), recently in Gupta et al. (2021), and has shown to be outperforming CNNs. A different approach is presented in Tong et al. (2020) using an audio–video cross-modal deep learning framework using both audio and video inputs. MFCCs and the corresponding deltas of the audio files and pre-processed frames of the video files are passed through independent CNN models to give feature cubes, which are then passed to a fully connected network for classification. Thus the advanced deep learning strategies raise the possibility of improving the efficient understanding of the spectral representation of speech.

1.2. Multi-head attention (MHA) mechanism and multi-task learning (MTL) approach

Introduced in Vaswani et al. (2017) for machine translation tasks, a ‘transformer’ is the transduction model that relies solely on the self-attention mechanism, without involving any recurrent or convolution operations. The self-attention or the intra-attention mechanism presents a weighted average of different feature representations, where the weights correspond to the relations between these representations. The MHA mechanism introduces parallel computations inside the self-attention. It has several attention layers running in parallel, which allows the model to jointly attend to information from different representation subspaces at different positions (Vaswani et al., 2017). They have proven to be efficient in the domain of natural language processing (NLP). When the MHA module provides cues to the main words in a sentence in NLP, it provides the main regions or important portions in an image to look at for interpretation. It is an idea rather than a module, to focus on areas containing key information. It can be interpreted as a vector of weights that give importance to different elements and indicates their correlation. It has been successfully used in image captioning (Xu et al., 2015), image classification (Dosovitskiy et al., 2020), speaker recognition (India et al., 2019) and speaker verification (India et al., 2021) tasks.

MTL is a training paradigm that works on the idea that, the machine learning models may benefit from information shared between different correlated tasks, when solved on the same data (Luu et al., 2020). MTL

is based on the learning process of human beings by which, they can execute multiple tasks accurately, with the integration of knowledge acquired on doing several tasks. This integration allows humans to rapidly learn with few examples, by learning concepts that are generalisable across multiple settings (Crawshaw, 2020). MTL allows the deep learning models to share common feature representations learned from multiple related tasks. This joint representation learning would improve data efficiency and can lead to faster learning for related tasks under data stringent conditions. Thus the high training data requirement and computational demands imposed by deep learning can be alleviated by MTL. On building deep learning models, this is typically implemented by sharing the initial layers while solving different tasks, and making the latter layers task-specific. Thus when the initial layers capture task-specific representation of the input data, this shared information is processed by the latter layers. MTL is actively used for computer vision (Liu et al., 2019), as well as for speech processing applications (Koizumi et al., 2020; Li et al., 2019; Tang et al., 2016).

1.3. Motivation

The dysarthric utterances have characteristics embedded in short segments like the bursts at the beginning and slurring periods or long silence regions in between, depending on their severity. Hence, we are motivated to use the attention mechanism to identify these salience periods from the spectrograms. Unlike natural images dealt with in computer vision domain, spectrograms are almost similar in appearance. When normal people find them totally similar, a trained speech specialist or a therapist can read from spectrograms with their 'experience'. This would help them to infer details present in the speech utterances represented by these spectrograms. This is because they know 'what' to look for, and 'where' to look at. Thus the underlying pathology or the severity characteristics can be intuitively identified by a deep learning model with the concept of attention. Attention mechanism has been able to improve the performance of a basic CNN network for thorax disease classification in Guan et al. (2018). The chest X-ray images where processed by the attention-guided CNN model to give the salient lesion regions to 'look at' for the classification procedure. We hypothesise that the attention mechanism could locate the salience periods from the spectrograms and could leverage the dysarthria severity recognition task.

The performance of automatic dysarthria severity assessment systems is strongly hindered due to the unavailability of large databases and the high intra-class variability. The former is due to the strain imposed on dysarthric patients for long recordings, and the latter is due to the differences in the type of dysarthria and associated health impairments of the patients within a particular class. MTL has leveraged the performance of speaker verification and diarization systems in Luu et al. (2020) by adding age and nationality as additional information. Nativity and gender information have similarly added advantages to speaker recognition in Montalvo et al. (2020). We hypothesise that the inherent differences in gender, age and the type of dysarthria can be learned jointly through MTL, and can mitigate the high intra-class variability in dysarthria severity estimation. Hence an MTL approach with these three auxiliary tasks is adopted. In literature, attention and MTL have been jointly implemented for various problems like speech enhancement (Koizumi et al., 2020), speech recognition (Qin et al., 2019), and speech emotion recognition (Li et al., 2019). When used in conjunction, they have shown to be efficient for these tasks, compared to the baseline works. Inspired by these results, we aim to do similarly using a CNN classifier with the spectrograms of the dysarthric utterances as input.

1.4. Contributions of the work

This work aims to use the time–frequency representation of dysarthric utterances via mel spectrograms with the advanced deep learning techniques introduced in the domain of computer vision. The major contributions of this work can be summed up as,

- Introduction of the MHA mechanism and the MTL approach for dysarthria severity level classification.
- Ablation study of the network to individually analysis the two methodologies.
- Analysis of the effectiveness of gender, age and disorder type identification as auxiliary tasks.
- Comprehensive evaluation on the Universal Access dysarthric speech corpus (UA-Speech) database, which allows comparison with the pioneer works in literature.

2. System description

The proposed framework for automated dysarthria severity classification is depicted in Fig. 1. The dysarthric utterances from the database are presented to the system in the form of mel spectrograms. The feature encoding power of CNN networks for image classification is exploited at the front-end. A ResCNN network is adopted for this, and it is appended with an MHA-based transformer block. At the final stage, the different auxiliary tasks are implemented to enable sharing of mutual information for conceptualising MTL. The system components are described in detail in the following subsections.

2.1. Front-end feature extraction

In this work, log mel spectrograms are used as the input features to represent the dysarthric utterances. The mel spectrograms mimic the human auditory system by smoothing the spectrograms to give high precision in the low frequencies and low precision in the high frequencies (O'shaughnessy, 1987). Thus it models the SLP's hearing perception to differentiate the dysarthria severity levels. Dysarthrics are found to have reduced vocal loudness, breathy/hoarse/harsh voice quality, reduced voice pitch inflections or monotone voice, and imprecise articulation (Dias et al., 2016). These varied auditory perceptual attributes are embedded in these spectrograms, and point to the underlying pathophysiology.

The UA-Speech database has utterances of duration varying between one second to 10 seconds in general. The vocal strain involved during speech production by the dysarthrics, and the pauses or breaks in speech segments result in longer duration of dysarthric utterances compared to their healthy counterparts. To preserve these relevant features, the variable length audio files are not clipped to constant duration, as in many of the previous works done in literature, but trimmed on both sides to remove the silence regions. Silence trimming is done using an energy-based voice activity detection at the beginning and end of the utterances alone, thus the pauses occurring in the voice segments are kept unchanged. Then overlapping triangular mel-scaled filters are employed to extract the log mel spectrograms. These spectrograms are preferred over the STFT-based spectrograms by the statistical classifiers (Chandrashekar et al., 2019b), and hence used here.

Fig. 2 shows the resized mel spectrograms of the word 'PSYCHOLOGICAL' corresponding to the utterances by speakers F03 (a 'HIGH' dysarthric) and F05 (a 'VERY LOW' dysarthric). The figures evidently show the reduction in the strength of formants and harmonics in the spectrogram of the 'HIGH' dysarthric, as shown by the second spectrogram, when compared to that of the 'VERY LOW' dysarthric. The poor articulation characteristics shown in the severe dysarthric utterance have led to reduced sharpness of the corresponding spectrogram. Thus, we can notice the efficiency of spectrograms in highlighting the

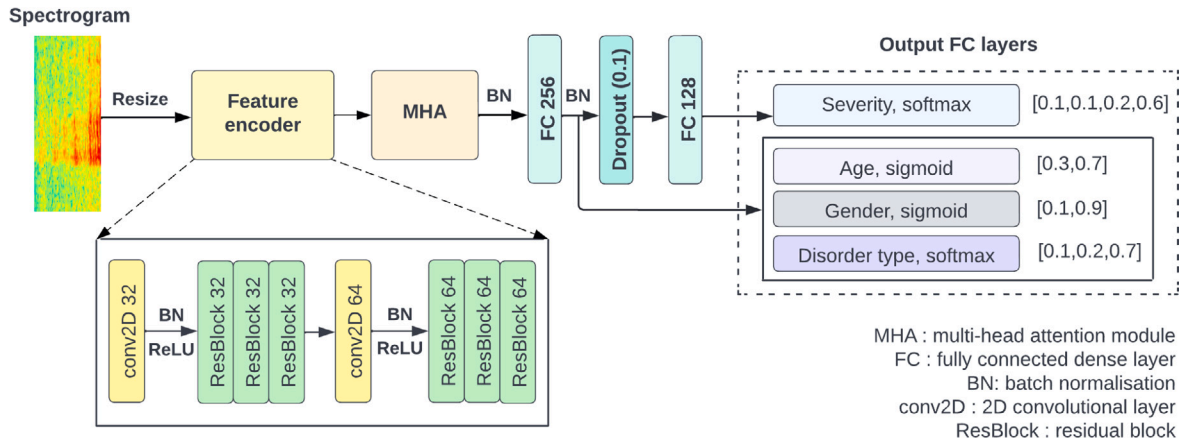

Fig. 1. Block diagram of the proposed system.

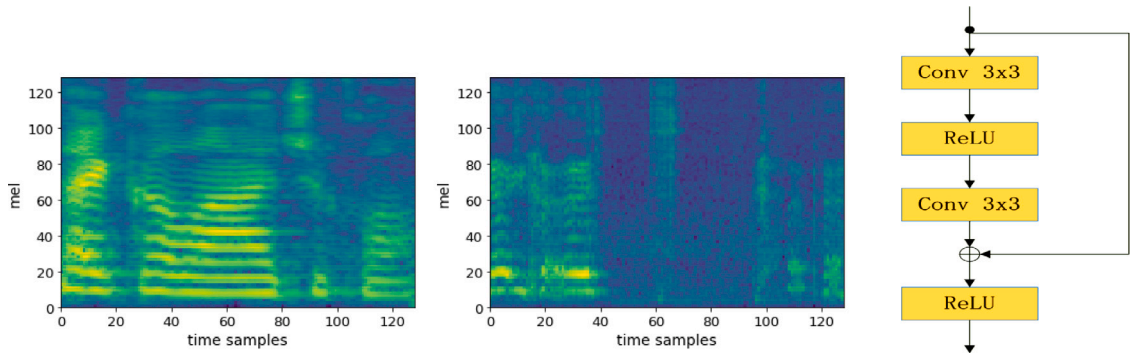

Fig. 2. Spectrograms of the utterance 'PSYCHOLOGICAL' spoken by the 'very low' dysarthric F05 (left) and the 'high' dysarthric F03 (middle); the ResBLock structure as reproduced from Li et al. (2017) showing the identity connection (right).

paralinguistic aspects through the spectro-temporal variations of the utterances. They are more intuitive in nature, compared to the one-dimensional speech characteristics. Hence, they prove to be capable of being discriminative features to be used with the deep learning classifiers. These spectrograms are then resized to 64 × 64 dimension. This is done because the CNN classifiers require fixed-size images for the input layer. The low dimension of 64 × 64 allows fast implementation, and can be altered in future studies.

2.2. Feature encoding

The mel spectrograms are fed to a CNN encoder which efficiently encodes the salient features. The encoder extracts the low-dimensional features from the spectrograms and efficiently represents the spatio-temporal information in the speech signals. The CNN feature extractor used is an adapted version of the ResCNN-based Deep Speaker architecture proposed for generating speaker embeddings in Li et al. (2017). Since dysarthria severity classification can be considered a speaker-group classification problem, we hypothesise that this model could be adapted suitably. The distinctive characteristics found in the spectrograms can be extracted by the ResCNN to distinguish the dysarthria severity levels. Fig. 2 shows the ResBLock structure at the right. The residual connections perform identity mapping with no additional parameters, and enable deeper networks to learn without over-fitting. It eases optimisation and introduces no additional computational complexity.

As in Li et al. (2017), the classifier has stacked up 2D convolutional (conv2D) layers of filter size 5 × 5 and stride 2 × 2, with the number of filters increasing from 32, in powers of 2. Each of these layers is followed by three ResBlocks, the number of filters in them being equal to the preceding conv2D layer. Deep Speaker was built for four conv2D-ResBlock structures, and was trained using triplet loss. However, our classifier is trained end-to-end by stochastic gradient descent (SGD) approach with back-propagation, and has only two blocks. Thus the number of blocks in the CNN feature map output is 16 × 16 and the number of feature maps at the end is 64.

2.3. MHA module

The attention mechanism was introduced in Bahdanau et al. (2014) for neural machine translation implemented using RNN-based encoder–decoder architectures. It tackled the bottleneck problem arising from the usage of fixed-length encoding vectors. This was done by bringing in the concept of using the most relevant portions of the input for decision-making. This flexible focusing is done using three attributes, namely query, values, and keys. The query is matched against a set of keys using a dot-product operation to generate a score value. These scores become the weights when passed through the softmax function. The value vectors corresponding to the keys are then weighed and summed to generate the 'attention'. The scaled dot-product attention and the MHA module comprising multiple parallel scaled dot-product attentions were proposed in Vaswani et al. (2017). From Vaswani et al. (2017), we have reproduced the visual representation of these modules in Fig. 3, where Q, V, and K represent matrices stacking the vectors query, values, and keys respectively. The query and key vectors are of size d_K and W^Q, W^K and W^V are the projection matrices used in generating the h different subspace representations of the query, key and value matrices. The assignment of the weight to each value is via a compatibility function of the query with the corresponding

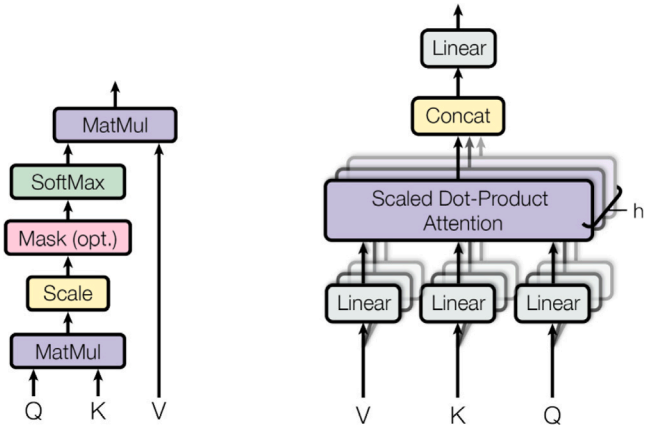

Fig. 3. Scaled dot-product attention (left) and multi-head attention (right).
Source: Reproduced from Vaswani et al. (2017).

key (Vaswani et al., 2017). The matrices Q, K and V are then given to generate the output of the attention function as,

$$attention(Q, K, V) = softmax(\frac{GV^T}{\sqrt{d_K}})V \quad (1)$$

Instead of performing a single attention function, MHA linearly projects the queries, keys, and values h times. The number of heads, h refers to the different learned projections. Upon each of these projections, a single attention mechanism is applied in parallel. These outputs are then concatenated and projected again to produce the final output. Thus information from different representation subspaces is obtained. MHA output can be obtained using the projection matrix W^O as,

$$MHA(Q, K, V) = concat(head_1, \ldots, head_h)W^O \quad (2)$$

where, each attention head is then computed as,

$$head_i = attention(QW_i^Q, KW_i^K, VW_i^V) \quad (3)$$

The attention mechanism can identify the different regions of interest from the encoded spectrograms. MHA is expected to capture different aspects from the same input, that can efficiently discriminate the severity levels. These can relate to the monotonicity, slurring effects, and long pauses found in the dysarthric utterances. It is observed that the utterances of the 'high' dysarthrics are of longer duration than that of 'low' dysarthrics, due to these pauses and nasal irregularities involved during the speech production, and the shivering and breaks in the voice segments. Identification of these features is relevant to the severity identification, and the objective function of the MHA module ensures that each head captures some dissimilar information.

Dimension of the linear space the input is to be projected after temporal summarisation is $64/h$. The output dimension of the MHA module after query-key vector multiplications and weighing of the value vectors is chosen to be 256. Thus we get a concatenation of the feature vectors generated by all the heads, and the network can extract different kinds of information from the different feature subspaces. This is followed by a normalisation layer for faster training and better convergence, and two fully connected dense layers of 256 and 128 units each. A batch normalisation layer and a 0.1 dropout layer are added in between these dense layers to improve the generalisation capability of the system. Output layers of the auxiliary tasks take the output of the dense layer with 256 units as input, whereas the main task takes the final dense layer output of 128 dimension.

2.4. MTL and classifier end

The UA-Speech database (Kim et al., 2008) used for evaluating the system has utterances from patients suffering from CP. These may exhibit characteristics such as slurred, slow, and less-intelligible speech with hoarse or breathy voice quality, depending on the severity and type of the disease. Based on the primary motor deficit, CP is neurologically classified as spastic, hypotonic, athetoid (dyskinetic), ataxic and mixed. Since the underlying pathomechanisms are different, there would be equivalent differences in the way the speech motor control is influenced, resulting in perceptually distinct dysarthria syndromes (Theresa Schölderle and Staiger, 2013). There are patients within the age group 18 to 58, and of both genders. Since age and gender are indexical variables in human speech, they introduce different linguistic patterns to the speech. Thus it is reasonable to check if there is any beneficial factor in these parameters for improving the dysarthria severity level estimation.

Table 1
Dysarthric speaker description of the UA-Speech database.

Severity	List of speakers	Age	Type
HIGH	M01, M04	< 30	Spastic
	M12	< 30	Mixed
	F03	>= 30	Spastic
MEDIUM	F02, M07, M16	>= 30	Spastic
LOW	F04	< 30	Athetoid
	M05	< 30	Spastic
	M11	>= 30	Athetoid
VERY LOW	F05, M08, M09	< 30	Spastic
	M10	< 30	Mixed
	M14	>= 30	Spastic

These auxiliary tasks are implemented by adding extra three dense layers at the end to generate the probability distributions over these tasks. The activation function of the output layer corresponding to age and gender is sigmoid, as they are binary classification problems, and thus the number of output nodes in these layers is two. The output layers corresponding to disorder type and severity level have three and four nodes, respectively, and are using softmax activation function. A sample output probability generation is depicted in the system block diagram in Fig. 1. The model is jointly optimised using the objective function:

$$L = \alpha L_{severity} + \beta L_{type} + \gamma L_{age} + \theta L_{gender} \quad (4)$$

where $L_{severity}$, L_{type}, L_{age} and L_{gender} are the losses for the classification tasks on dysarthria severity level, the disorder type, the speaker age, and the gender, respectively. α, β, γ and θ are their respective loss weights. The value of α always remains one, while the other weights are varied between 0.25 and 1 to analyse the impact of the auxiliary tasks on the severity classification. The classification of speakers based on the above-mentioned auxiliary tasks can be viewed in the different columns of Table 1.

3. Database

The UA-Speech database (Kim et al., 2008) is used for evaluating the proposed system. It has utterances from 13 healthy speakers and 19 dysarthrics diagnosed with CP. However, the data of only 15 patients are available, as given in Table 1. The severity levels are assigned as — very low, low, medium and high, based on the intelligibility ratings by five naive listeners, as follows: (0–25)%-high, (25–50)%-medium, (50–75)%-low and (75–100)%-very low. The first letter of the speaker's name 'F' or 'M' indicates the gender, and thus there are four female speakers and 11 male speakers. The speakers are categorised based on their age into two groups: aged 30 and above, or aged below 30. They are also differentiated based on their disorder type as indicated in Table 1. The number of speakers per class would seem to be small. But the UA-Speech database is the latest and the largest of the available dysarthric speech databases with speakers from all four severity levels. The 15 dysarthric subjects evaluated contribute to about 17 hours of

speech. This is because speaking is a tiring task for the dysarthrics and their physical fatigue and frustration hinder long recordings.

The utterances present in the database correspond to three repetitions of the 10 English digits, 19 computer commands, 26 international radio alphabets, 100 common words in the Brown corpus and 300 distinct uncommon words, totalling 765 word utterances per speaker. We used the common words alone for training, and the uncommon words for testing, which means that the network is evaluated on unseen words, which measures its robustness. Thus we get 465 words per speaker for training and 300 per speaker for testing. The audio files are sampled at f_s = 16 kHz and is recorded through an 8-microphone array, as well as with a digital video camera (Kim et al., 2008). The sixth channel in the array had the highest signal-to-noise ratio, and hence those audio files are used in the work.

4. Experimental framework

4.1. Baseline system

Being the recent work reporting high accuracy for the proposed task, we adopt the best method of Chandrashekar et al. (2020) as the baseline system. CQT spectrograms are used with a time–frequency CNN model for the classification of the dysarthria severity levels. The authors had analysed the effect of varying minimum frequency (f_{min}) and the number of bins per octave (b). They have concluded that 60 Hz and 120 Hz work best as f_{min} for male and female speakers respectively, as the fundamental frequency is lesser for the male speakers compared to the female speakers. b was set to 12, and increased in steps of 12 to 48. It was observed that the frequency resolution improved with such an increase, while the temporal resolution decreased. We used 24 as b, as it had the best accuracy. Thus, the CQT spectrograms are extracted with these parameters and the maximum frequency ($f_{max} = f_s/2$) as 8 kHz.

The time–frequency CNN architecture employed in the work had one input layer, three hidden layers, and a fully connected layer of softmax activation with hidden units equal to the number of classes. We have built the CNN classifier with the same specification: convolution layer followed by max-pooling in the first two hidden layers, and a convolutional layer alone in the third one. All the layers have 32 filters of kernel size 3 × 3. They have no padding, a stride of 2 and ReLu activation. The input dimension is 100 × 100 × 3. The model was trained and validated for five epochs. Each epoch had 17 iterations and the learning rate was chosen to be 0.001.

4.2. Feature design and model evaluation

After silence trimming at both ends, the variable length log mel spectrograms are extracted using 128 overlapping triangular mel scaled filters, for a frame size of 25 ms and frame shift of 5 ms. These are then resized to 64 × 64 dimension. The models are evaluated by training them for 60 epochs, with early stopping applied after a patience of 30 epochs. 10% of the training data is used for validation. The batch-size is chosen to be 128 (best among 32, 64 and 128) and the learning rate to be 0.001 (best among 0.01, 0.001 and 0.0001), after hyper-parameter tuning on the validation data. Categorical cross entropy and softmax act as the loss function and activation function respectively for $L_{severity}$ and L_{type}, while binary cross entropy and sigmoid function act alike for L_{age} and L_{gender}. As mentioned, SGD is the optimiser used with a momentum of 0.9 (best among SGD, Adam and adaptive delta).

5. Results and analysis

Performance analysis of the proposed system against the baseline system, and the detailed study on varying the number of attention heads and loss weights are explained below.

Table 2
Variation of the severity classification accuracy (%) with the number of attention heads and loss weights of the auxiliary tasks (the best result in bold).

Method	Parameter	Accuracy			
MHA	Heads	1	2	4	8
	h	82.67	84.27	**87.49**	86.15
MTL	Loss weights	0.25	0.5	0.75	1
	β	59.42	**91.20**	90.69	89.28
	γ	63.55	47.89	**67.55**	33.33
	θ	89.31	89.31	90.09	**90.75**

5.1. Impact of number of attention heads

MHA is more expressive than vanilla attention models, and the system performance is influenced by the number of parallel attention layers, or heads (h). With the increasing number of heads, the accuracy is likely to improve. It is a hyper-parameter to be looked at for improving the system performance. It is varied between 1 to 8, and the results are shown in the first row of Table 2. This reports the performance of the ResCNN+MHA system, without including the MTL approach. The best accuracy is found at h = 4, and as expected the accuracy drops off with too many heads (Vaswani et al., 2017). With the increasing number of heads, the number of trainable parameters increases. Also, since the heads within a layer are independent, in some cases, some of the heads may be capturing irrelevant or less useful data. It has also been shown in Michel et al. (2019) that, many of the attention heads can be removed at test time without significantly impacting the system performance. Thus the number of heads considered in our study is not further increased.

5.2. Impact of loss weights

The MTL approach was incorporated by adding extra dense layers at the end for the auxiliary task classification. By this concept of 'hard parameter sharing', the model would be less prone to over-fitting. But to identify the impact of each of the adopted auxiliary tasks, the loss weights are tuned. The effect of varying the loss weights of the auxiliary tasks is shown in the latter rows of Table 2, on the ResCNN+MTL system, without MHA module. The weights of the different tasks are changed from 0.25 to 1 to study their individual contribution to the main task. This was done by taking each task alone as an auxiliary task to the severity identification. Hence, the ResCNN+MTL system has two softmax dense layers at the end, one for the main dysarthria classification, and another for the auxiliary task whose weight is being tuned. The best value of θ being one indicates that, incorporating gender identification is really beneficial to the severity classification. It has mitigated the confusion differentiating the male and female dysarthric speech and improved the baseline model accuracy by over 4%. However, the best accuracy on using type identification alone as the auxiliary task was obtained for β = 0.5, and a further increase led to deterioration. This shows that overweighing the auxiliary task will lead to insignificant feature learning for the severity identification, at the gain of improved disorder-type classification accuracy. Thus, a proper choice of loss weights is important in MTL. Considering the age group classification alone as an auxiliary task has led to a reduction in the accuracy of the model for all values of γ. This demonstrates the phenomenon of negative transfer or destructive interference, wherein the performance of the main task decreases as the model focuses more on improving the results of the auxiliary task. The features shared would not be beneficial and thus a proper choice of the auxiliary task is also very important in MTL.

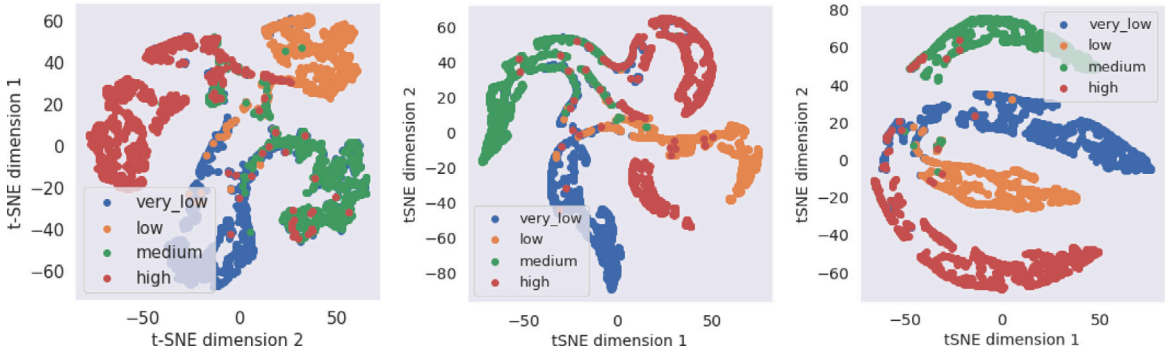

Fig. 4. t-SNE plots of the baseline model (left), ResCNN model (middle) and the proposed model (right).

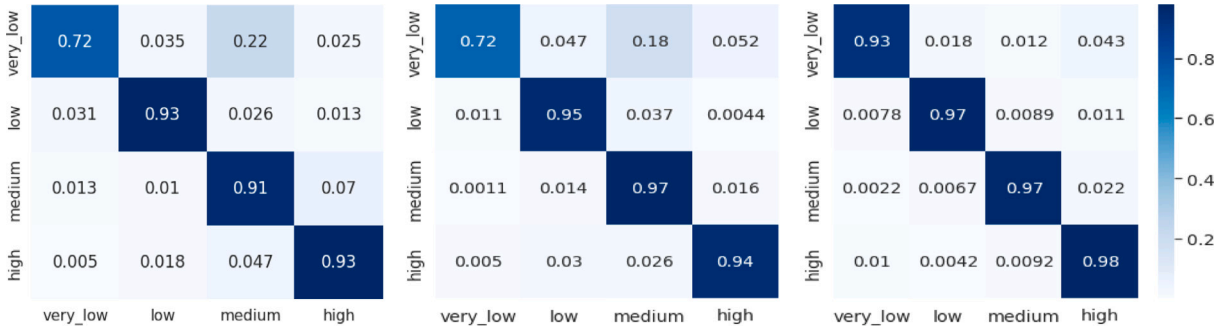

Fig. 5. Confusion matrix given by the baseline model (left), ResCNN model (middle) and the proposed model (right).

Table 3
Severity classification accuracy of the different classifiers (%) (the best result in bold).

SI no.	Classifier	Accuracy
1	CQT-CNN (Chandrashekar et al., 2020)	84.24
2	ResCNN	87.14
3	ResCNN + MHA	87.49
4	ResCNN + MTL	91.11
5	ResCNN + MTL2	92.02
6	**ResCNN + MHA + MTL2**	**95.75**

5.3. Ablation study

Table 3 shows the performance of the various systems against the baseline system. For the independent implementation of the ResCNN model, after the conv2D-ResBlocks, the affine and normalisation layers are added. For the ResCNN+MHA system, the MHA module follows the ResCNN encoding block, and the ResCNN+MTL system has four dense layers instead of one, with the best-chosen loss weight values, namely, $\beta = 0.5$, $\gamma = 0.75$ and $\theta = 1$. To mitigate the effect of negative transfer, MTL was performed excluding age identification, giving the ResCNN+MTL2 model. A better result was obtained and this means that, there is no significant difference in the severity characteristics with the patient's age, and adding age-related features would lead to misclassifications of the severity level. Hence, the final model is implemented using the same strategy and is referred to as ResCNN+MHA+MTL2.

An accuracy of 92.76% was initially obtained under this setting, but to get a better-refined model, a grid search was done for different values of β and θ on this final model. Thus on the ResCNN model with the MHA module, a final round of hyper-parameter tuning was done with respect to β and θ, which correspond to the advantageous auxiliary tasks of disorder-type and gender identifications. The best classification accuracy of 95.75% was obtained for $\beta = 0.75$ and $\theta = 1$. Hence, these loss weights are chosen for the proposed model.

In general, the more the number of related tasks the model learns simultaneously, the less its chance of overfitting. This is because, by learning multiple tasks, it identifies a representation capturing the important factors contributing to all of the tasks and hence, less prone to overfitting on the main task. But the ablation study demonstrates the need to properly identify the auxiliary tasks and their corresponding loss weights, so as to boost the model's performance on the main task.

It can be observed that the proposed system leads the baseline system by over 11%, and the ResCNN model by over 8%, which suggests that both MHA and MTL substantially contribute to the severity classification accuracy. This is again validated by the t-distributed stochastic neighbour embedding (t-SNE) (Van der Maaten and Hinton, 2008) plots drawn from the output vectors produced by the snippets from the last dense layer of the trained models in Fig. 4. Good clustering is exhibited by the proposed model, in contrast to that shown by the baseline models. Thus the severity levels are well differentiated by the proposed system. The normalised confusion matrices of the baseline models and the proposed system are shown in Fig. 5. It can be observed that, a minimum of 93% accuracy is guaranteed in the identification of all the four severity levels by the proposed system, in contrast to the baseline systems which work poorly on the border class 'very low'.

5.4. Statistical analysis

The proposed model was evaluated by testing on 'uncommon words' of the UA-Speech database, after being trained on all the other utterances. This train–test partition is the widely used data handling strategy on the UA-Speech database by almost all the pioneer works in literature, some of them being Gurugubelli and Vuppala (2019), Tripathi et al. (2020b,a), Chandrashekar et al. (2019b), Joshy and Rajan (2021, 2022). But since the available training data is limited, and the results are not reported using any cross-validation strategy, the improvement of 11.51% obtained over the baseline may be doubted to be unreal, as a result of any statistical fluke. To reaffirm the importance of the proposed approach, the classifiers are compared using the widely used statistical test, namely, McNemar's statistical hypothesis

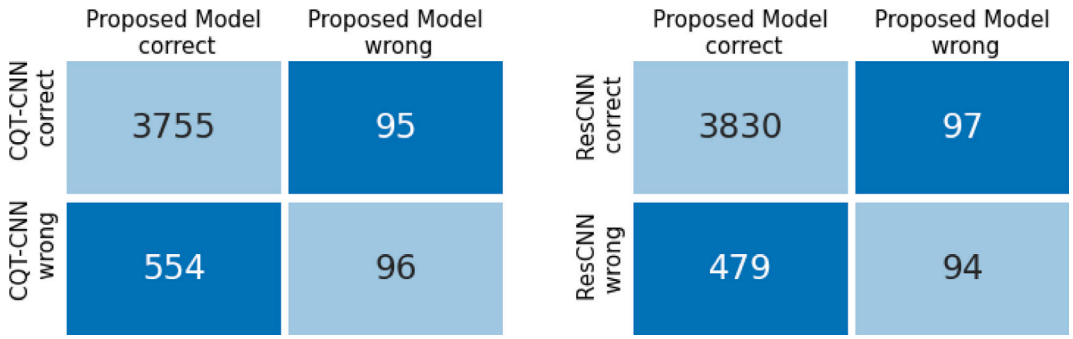

Fig. 6. Contingency tables given by the proposed model against the CQT-CNN model (left) and the ResCNN model (right).

test (Everitt, 1992). This test was adopted as per the findings of Dietterich in Dietterich (1998) since our test data is fixed, comprising of the uncommon words of every speaker, and repeated test cases are hence not needed. This is different to the random/fixed subsets used in repeated evaluations as in the ensemble/k-fold cross-validation methods using any resampling technique. Hence, the data variability issue occurring due to the selection of train–test split in the McNemar test results is guaranteed to be absent in our test case.

The skill measure adopted for comparing the models is the classification accuracy. The contingency table is constructed based on the success(1)/failure(0) measure of the two models being compared. It is of the form,

$$\begin{bmatrix} n11 & n01 \\ n10 & n00 \end{bmatrix}$$

where, $n11$ indicates the count of the dysarthric utterances that were correctly classified by both the models, and $n10$ indicates the count of the utterances correctly classified by model 1 but misclassified by model 2. Similarly the other two counts $n01$ and $n00$ are defined. Thus the total number of samples in the test set would be the sum of these, as $n = n00 + n01 + n10 + n11$. When doing the statistical hypothesis test, the null hypothesis ($H0$) is defined as the condition $n01 = n10$, that is the two models have the same error rate or the same proportion of misclassifications. The McNemar's test checks for the marginal homogeneity in the contingency table by testing if there is a significant difference between the counts $n01$ and $n10$. This is done using the test statistic t, defined in Everitt (1992) to include the continuity correction term -1 in the numerator as,

$$t = \frac{(|n01 - n10| - 1)^2}{(n01 + n10)} \tag{5}$$

This test statistic (t statistic) has a Chi-Squared distribution with 1 degree of freedom, and if $H0$ is accepted, then the probability that $t > \chi^2_{1,0.95} = 3.841459$ is less than $\alpha = 0.05$. This test is implemented in Python using the mcnemar() function of the Statsmodels module.

The p-value calculated from t statistics is compared with an alpha value to make the final decision as

- $p > \alpha$: fail to reject $H0$, both models have a similar proportion of errors on the test dataset.
- $p \leq \alpha$: reject $H0$, there is a significant difference in the proportion of errors, indicating one is better than the other.

The contingency tables obtained from the McNemar test done on the proposed ResCNN+MHA+MTL2 model against the baseline CQT-CNN model and the basic ResCNN model are shown in the left and right figures in Fig. 6 respectively. We can find the difference in the proportions of the errors by looking at the values corresponding to $n01$ and $n10$. A large difference is clearly visible, which indicates the effectiveness of using the proposed model against the baseline systems. On calculating the test statistics, $t = 323.21$ and $t = 252.02$ were obtained respectively, which both resulted in 0.00 p-values. Hence $H0$ is rejected in both cases on taking $\alpha = 0.05$, which proves that the margins of accuracy score gained by the proposed system are statistically significant.

Further analysis was done using the precision (P), recall (R), F1 score, and the area under the ROC curve (AUC) measures, to compare the proposed model against the baseline systems. These measures are significant since the dataset studied is of limited size and is unbalanced in the number of speakers within the four severity levels. Table 4 gives the results. We find that when the ResCNN model was capable of lifting the performance of the baseline CQT-CNN model in terms of almost all measures, it could not improve the recall rate and F1 scores of the 'very low' class significantly. However, the proposed model was capable of this. Also, we find a large difference in the P(0.67 to 0.95 by CQT-CNN and 0.72–0.98 by ResCNN) and R(0.72 to 0.93 by CQT-CNN and 0.72–0.97 by ResCNN) values among the different classes for the two baseline models compared to the almost uniform performance over various classes given by the proposed model (0.93 to 0.99 for P and 0.93 to 0.98 for R).

5.5. Speaker-dependency check

To evaluate the efficacy of the proposed system on unseen speakers, a comprehensive round-robin leave-one-speaker-out (LOSO) experiment is performed. This is the most applied approach for analysing the system's performance under the SID scenario in the pioneer works on dysarthric severity classification such as Gurugubelli and Vuppala (2019), Tripathi et al. (2020b), Chandrashekar et al. (2019b), Joshy and Rajan (2022). Since there are 15 speakers in the UA-Speech database, 15 rounds of experimentation are required, whose average is taken. This means that, in each round, data from 14 speakers would be taken for training, and the data of the left-out speaker is taken for testing. The test is done using the seen (465 words used in training) and unseen (300 uncommon words preserved for testing) words. These are referred to as 'Test 1' and 'Test 2' respectively. Thus when Test 1 measures the system robustness against a new dysarthric speaker, Test 2 checks for its robustness against new vocabulary as well. These are important in understanding the system's applicability in evaluating a new dysarthric speaker.

Table 5 gives the results of the LOSO cross-validation experiment on the proposed model and the baseline classifiers. We find that there is a good margin of improvement in terms of the average classification accuracy over the baseline classifiers by the proposed model on both tests. All the classifiers give acceptable results on the border classes 'very low' and 'high' under the SID scenario but very poor results on the intermediate classes 'low' and 'medium'. This is in agreement with the findings reported in Tripathi et al. (2020b), Joshy and Rajan (2022) and occurs due to the unbalanced nature of the UA-Speech database. The baseline CQT-CNN classifier has almost 0% accuracy in dealing with these classes, as they were mapped to the nearby border classes comprising more speakers during training.

When the proposed model was implemented with the best performing loss weights in the SD scenario, namely $\beta = 0.75$ and $\theta = 1$, the

Table 4
Precision (P), recall (R), F1 score and area under the ROC curve (AUC) measures of the different classifiers.

Severity level	CQT-CNN				ResCNN				Proposed model			
	P	R	F1	AUC	P	R	F1	AUC	P	R	F1	AUC
Very low	0.95	0.72	0.82	0.85	0.98	0.72	0.83	0.86	0.99	0.93	0.95	0.96
Low	0.91	0.93	0.91	0.95	0.88	0.95	0.91	0.96	0.96	0.97	0.97	0.98
Medium	0.67	0.91	0.77	0.90	0.72	0.97	0.83	0.94	0.96	0.97	0.96	0.98
High	0.91	0.93	0.91	0.95	0.92	0.94	0.93	0.95	0.93	0.98	0.95	0.97

Table 5
Average LOSO cross-validation accuracy (in %) for the classifiers using known words (Test 1) and unknown words (Test 2) for testing.

Severity level	Test 1			Test 2		
	CQT-CNN	ResCNN	Proposed model	CQT-CNN	ResCNN	Proposed model
Very low	52.13	51.01	64.52	53.80	36.53	47.80
Low	1.22	6.67	16.27	0.33	7.77	15.55
Medium	0.51	10.03	11.90	0.88	16.22	21.22
High	20.59	42.47	42.42	21.99	41.83	46.25
Total	23.21	31.67	**38.45**	24.04	28.13	**35.62**

average accuracy was only 34.69% for Test 1 and 30.31% for Test 2. We investigated if the conflicting gradients of the different tasks prevented the trunk (lower-level layers with shared representations) from fully utilising the different task-specific information to improve the dysarthria severity classification. So, the weights of the auxiliary tasks were tuned again with smaller values [0.15, 0.25, 0.35, 0.5], and the best performing classifier was found to be using $\beta = 0.25$ and $\theta = 0.35$. The results of this model are reported in Table 5, and an appreciable gain over the baseline classifiers can be observed in all the severity levels. As expected, the accuracy of correctly identifying the dysarthric severity level of a new speaker from the words seen during training (Test 1) is higher than the accuracy observed on using unseen words (Test 2). This again points to the fact that the speech deficits common to the speakers of each severity level are visibly seen in their utterances of the same word, such as a pause in between, or a missed/repeated phoneme.

The proposed model gave an accuracy of 35.62% on testing with uncommon words, against the best accuracy of 54% reported in the literature Tripathi et al. (2020b) under the SID scenario with a similar testing strategy. But it is important to note that this gain was obtained by using the posteriors of the DeepSpeech-1 ASR model that used 1000 hours of data during training, and the system was specifically built for the SID setting. In this work, we aimed to investigate if the chosen auxiliary tasks synergise to understand the underlying dysarthric characteristics to properly identify the severity levels, under the stringent data conditions. It was found from the SID results that, even in the intermediate classes the proposed approach could uplift the performance of the baseline ResCNN system. This insight can be further developed to build a reliable dysarthria severity assessment system. Improvement can be further obtained using advanced loss weighting strategies such as dynamic weight average and uncertainty weighting.

5.6. Discussion

To the best of our knowledge, this is the first detailed investigation on MHA and MTL for dysarthria severity classification. We obtained a classification accuracy of 95.75%, at a gain of 11.51% over the baseline system, which is a good margin. This throws light on using MHA and MTL for improving the performance of the proposed task. It was found that the age of the dysarthric patient does not correlate with the severity characteristics, and hence cannot map to the correct severity level of the patient. Thus the study also brings out the fact that proper auxiliary tasks have to be chosen for implementing MTL. In comparison with the pioneer works in literature, we found that the obtained result is appreciable. The results of Chandrashekar et al. (2019b, 2020) a nd Gupta et al. (2021) are reported on subsets of the UA-Speech database and hence not comparable. Tripathi et al. (2020b) reports an accuracy of 97.40% using SID textual-derived features extracted using a pre-trained Deep Speech-1 ASR model, trained with 1000 hours of data. However, our result was obtained on approximately 17 hours of dysarthric speech, and highlights the efficacy of MHA and MTL approaches for the proposed task even with limited data. An improvement is surely expected with the availability of larger data. However, the difficulty in speaking faced by the dysarthrics leads to difficulty in data collection, resulting in the low resource of dysarthric speech data. This is the main reason for the low accuracy obtained under the SID scenario. Even then, the model proved to be more efficient than the baseline system by giving a margin of 11.58%.

Further improvement in the proposed system performance can potentially be achieved by using a better time–frequency representation, and an advanced fine-tuned feature encoder at the front end. We would like to explore the potency of Gabor spectrograms at the feature-end as future work. The residual network can possibly be improved by using the squeeze-excitation (SE) mechanism or the more recent competitive SE mechanism with inner-imaging. This would reweight the channel-wise responses and model their inter-dependencies using the self-attention mechanism. Also, an analysis of building architectures for the task-specific branches across the three different tasks can be done on the improved shared trunk. Data augmentation also can be experimented as future work, with a deep convolutional generative adversarial network (DCGAN) or a wave generative adversarial network (WaveGAN). DCGAN can generate additional mel spectrograms for the training procedure, whereas WaveGAN can generate audio files, which can be listened for manual inspection. Again, the data scarcity would pose a challenge to such a system.

6. Conclusion

Automated assessment of the dysarthria severity level can help clinicians for easy diagnosis of the progression of the disease. This work analysed the effect of incorporating an MHA module and adopting the MTL approach at the classifier end for dysarthria severity level classification. Joint learning of the different subspace representations by the former mechanism, and shared feature descriptions about the gender and disorder type by the latter were found to be beneficial in correctly detecting the severity level. This novel approach promises an enhancement in the performance of an automated dysarthria severity classification system under data-stringent conditions.

CRediT authorship contribution statement

Amlu Anna Joshy: Conception and design of study, Acquisition of data, Analysis and/or interpretation of data, Writing – original draft, Writing – review & editing. **Rajeev Rajan:** Conception and design of study, Acquisition of data, Analysis and/or interpretation of data, Writing – original draft, Writing – review & editing.

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.

Data availability

The authors do not have permission to share data.

Acknowledgement

All authors approved the version of the manuscript to be published.

References

An, K., Kim, M.J., Teplansky, K., Green, J.R., Campbell, T.F., Yunusova, Y., Heitzman, D., Wang, J., 2018. Automatic early detection of amyotrophic lateral sclerosis from intelligible speech using convolutional neural networks. In: Proc. Interspeech. pp. 1913–1917.

Bahdanau, D., Cho, K., Bengio, Y., 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

Bhat, C., Strik, H., 2020. Automatic assessment of sentence-level dysarthria intelligibility using BLSTM. IEEE J. Sel. Top. Signal Process. 14, 322–330.

Bhat, C., Vachhani, B., Kopparapu, S.K., 2017. Automatic assessment of dysarthria severity level using audio descriptors. In: Proc. IEEE Int. Conf. Acoust. Speech Signal Process.. pp. 5070–5074.

Chandrashekar, H.M., Karjigi, V., Sreedevi, N., 2019a. Breathiness indices for classification of dysarthria based on type and speech intelligibility. In: Proc. IEEE Int. Conf. Wireless Commun. Signal Process. Network. pp. 266–270.

Chandrashekar, H.M., Karjigi, V., Sreedevi, N., 2019b. Spectro-temporal representation of speech for intelligibility assessment of dysarthria. IEEE J. Sel. Top. Signal Process. 14, 390–399.

Chandrashekar, H.M., Karjigi, V., Sreedevi, N., 2020. Investigation of different time-frequency representations for intelligibility assessment of dysarthric speech. IEEE Trans. Neural Sys. Rehab. Engn. 2880–2889.

Crawshaw, M., 2020. Multi-task learning with deep neural networks: A survey. arXiv preprint arXiv:2009.09796.

Dias, A.E., Barbosa, M.T., Limongi, J.C.P., Barbosa, E.R., 2016. Speech disorders did not correlate with age at onset of Parkinson's disease. Arquivos Neuro-Psiquiatria 74, 117–121.

Dietterich, T.G., 1998. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput. 10, 1895–1923.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16 × 16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Drummond, S.S., 1993. Dysarthria Examination Battery–Manual/instructions for Administration. Communication Skill Builders, Tuscon (Arizona).

Enderby, P., 1980. Frenchay dysarthria assessment. Brit. J. Disorders Commun. 15, 165–173.

Everitt, B.S., 1992. The Analysis of Contingency Tables. CRC Press.

Guan, Q., Huang, Y., Zhong, Z., Zheng, Z., Zheng, L., Yang, Y., 2018. Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification. arXiv preprint arXiv:1801.09927.

Gupta, S., Patil, A.T., Purohit, M., Parmar, M., Patel, M., Patil, H.A., Guido, R.C., 2021. Residual neural network precisely quantifies dysarthria severity-level based on short-duration speech segments. Neural Netw. 105–117.

Gurugubelli, K., Vuppala, A.K., 2019. Perceptually enhanced single frequency filtering for dysarthric speech detection and intelligibility assessment. In: Proc. IEEE Int. Conf. Acoust. Speech Signal Process.. pp. 3403–3407.

India, M., Safari, P., Hernado, J., 2021. Double multi-head attention for speaker verification. In: Proc. IEEE Int. Conf. Acoust. Speech Signal Process.. IEEE, pp. 6144–6148.

India, M., Safari, P., Hernando, J., 2019. Self multi-head attention for speaker recognition. arXiv preprint arXiv:1906.09890.

Joshy, A.A., Rajan, R., 2021. Automated dysarthria severity classification using deep learning frameworks. In: Proc. 28th Eur. Signal Process. Conf.. pp. 116–120.

Joshy, A.A., Rajan, R., 2022. Automated dysarthria severity classification: A study on acoustic features and deep learning techniques. IEEE Trans. Neural Syst. Rehabil. Eng. 30, 1147–1157.

Kadi, K.L., Selouani, S.A., Boudraa, B., Boudraa, M., 2013. Discriminative prosodic features to assess the dysarthria severity levels. In: Proc. World Congress on Engg., vol. 3.

Kadi, K.L., Selouani, S.A., Boudraa, B., Boudraa, M., 2016. Fully automated speaker identification and intelligibility assessment in dysarthria disease using auditory knowledge. Biocybern. Biomed. Eng. 36, 233–247.

Kent, R.D., Weismer, G., Kent, J.F., Vorperian, H.K., Duffy, J.R., 1999. Acoustic studies of dysarthric speech: Methods, progress, and potential. J. Commun. Disord. 32, 141–186.

Kim, H., Hasegawa Johnson, M., Perlman, A., Gunderson, J., Huang, T.S., Watkin, K., Frame, S., 2008. Dysarthric speech database for universal access research. In: Ninth Annual Conf. Int. Speech Commun. Asso.. pp. 1741–1744.

Koizumi, Y., Yatabe, K., Delcroix, M., Masuyama, Y., Takeuchi, D., 2020. Speech enhancement using self-adaptation and multi-head self-attention. In: Proc. IEEE Int. Conf. Acoust. Speech Signal Process.. IEEE, pp. 181–185.

Li, C., Ma, X., Jiang, B., Li, X., Zhang, X., Liu, X., Cao, Y., Kannan, A., Zhu, Z., 2017. Deep speaker: An end-to-end neural speaker embedding system. 650, arXiv preprint arXiv:1705.02304.

Li, Y., Zhao, T., Kawahara, T., 2019. Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning.. In: Proc. Interspeech. pp. 2803–2807.

Liu, S., Johns, E., Davison, A.J., 2019. End-to-end multi-task learning with attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1871–1880.

Luu, C., Bell, P., Renals, S., 2020. Leveraging speaker attribute information using multi task learning for speaker verification and diarization. arXiv preprint arXiv:2010.14269.

Van der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE. J. Machine Learning Research 9.

Martínez, D., Lleida, E., Green, P., Christensen, H., Ortega, A., Miguel, A., 2015. Intelligibility assessment and speech recognizer word accuracy rate prediction for dysarthric speakers in a factor analysis subspace. ACM Trans. Access. Comput. 6, 1–21.

Michel, P., Levy, O., Neubig, G., 2019. Are sixteen heads really better than one? Adv. Neural Inf. Process. Syst. 32.

Millet, J., Zeghidour, N., 2019. Learning to detect dysarthria from raw speech. In: Proc. IEEE Int. Conf. Acoust. Speech Signal Process.. IEEE, pp. 5831–5835.

Montalvo, A., Calvo, J.R., Bonastre, J.-F., 2020. Multi-task learning for voice related recognition tasks. In: Proc. Interspeech. pp. 2997–3001.

O'shaughnessy, D., 1987. Speech Communications: Human and Machine (IEEE). Universities Press.

Prabhakera, N., Alku, P., 2018. Dysarthric speech classification using glottal features computed from non-words, words and sentences. In: Proc. Interspeech. pp. 3403–3407.

Qin, C.-X., Zhang, W.-L., Qu, D., 2019. A new joint CTC-attention-based speech recognition model with multi-level multi-head attention. EURASIP J. Audio Speech Music Process. 2019, 1–12.

Qualls, C.D., Battle, D.E., 2012. Neurogenic disorders of speech language cognition-communication and swallowing. In: Communication Disorders in Multicultural and International Populations. Mosby, pp. 148–163.

Robertson, S.J., 1982. Robertson Dysarthria Profile. Buckinghamshire: Winslow.

Rudzicz, F., 2010. Articulatory knowledge in the recognition of dysarthric speech. IEEE Trans. Audio, Speech Lang. Process. 19, 947–960.

Schlenck, K.J., Bettrich, R., Willmes, K., 1993. Aspects of disturbed prosody in dysarthria. Clin. Linguist. Phon. 7, 119–128.

Shriberg, L.D., Kwiatkowski, J., 1982. Phonological disorders III: A procedure for assessing severity of involvement. J. Speech Hear. Disord. 47, 256–270.

Suhas, B., Mallela, J., Illa, A., Yamini, B., Atchayaram, N., Yadav, R., Gope, D., Ghosh, P.K., 2020. Speech task based automatic classification of ALS and Parkinson's disease and their severity using log mel spectrograms. In: Int. Conf. Signal Process. Comm.. IEEE, pp. 1–5.

Tang, Z., Li, L., Wang, D., 2016. Multi-task recurrent model for speech and speaker recognition. In: 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. IEEE, pp. 1–4.

Theresa Schölderle, M.A., Staiger, A., 2013. Dysarthria syndromes in adult cerebral palsy. J. Med. Speech-Lang. Pathol. 20, 100–105.

Tong, H., Sharifzadeh, H., McLoughlin, I., 2020. Automatic assessment of dysarthric severity level using audio-video cross-modal approach in deep learning. In: Proc. Interspeech. pp. 4786–4790.

Tripathi, A., Bhosale, S., Kopparapu, S.K., 2020a. A novel approach for intelligibility assessment in dysarthric subjects. In: Proc. IEEE Int. Conf. Acoust. Speech Signal Process.. pp. 6779–6783.

Tripathi, A., Bhosale, S., Kopparapu, S.K., 2020b. Improved speaker independent dysarthria intelligibility classification using deepspeech posteriors. In: Proc. IEEE Int. Conf. Acoust. Speech Signal Process.. pp. 6114–6118.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 5998–6008.

Vyas, G., Dutta, M.K., Prinosil, J., Harár, P., 2016. An automatic diagnosis and assessment of dysarthric speech using speech disorder specific prosodic features. In: Proc. IEEE Int. Conf. Telecommun. Signal Process.. pp. 515–518.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y., 2015. Show, attend and tell: Neural image caption generation with visual attention. In: International Conference on Machine Learning. PMLR, pp. 2048–2057.