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

Amlu Anna Joshy, Rajeev Rajan, (2023)

Presented by: Emilie Tortel, Nora Raud

Aalto University
School of Electrical
Engineering

Table of Contents

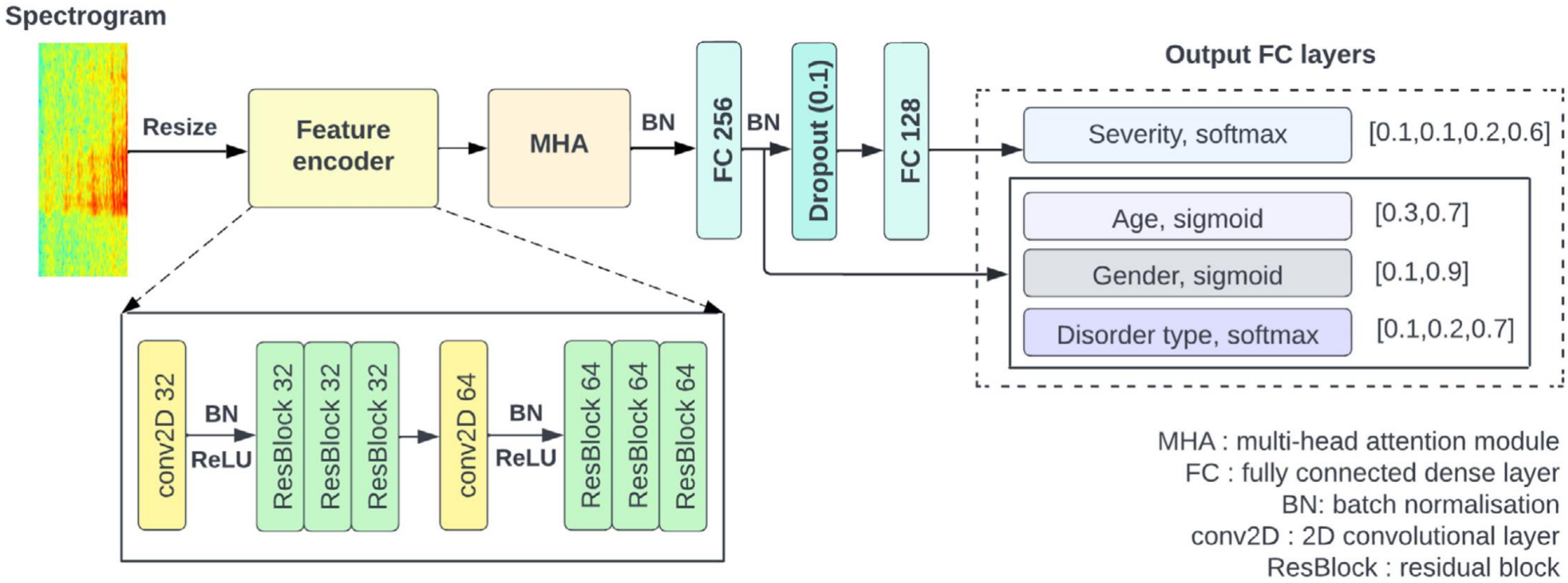
Motivation

- Slow down **progressive dysarthria**
- Economical, consistent
- Facilitate communication for people with dysarthria

Research Gap

- Previous approaches:
  - Prosodic features as input
  - Time-frequency representation as input
  - ANNs and CNNs
  - **Baseline model**: CQT-CNN
- Need for an efficient understanding of the **spectral representation** of speech
- Novel approach :
  - Multi-head attention
  - Multi-task learning

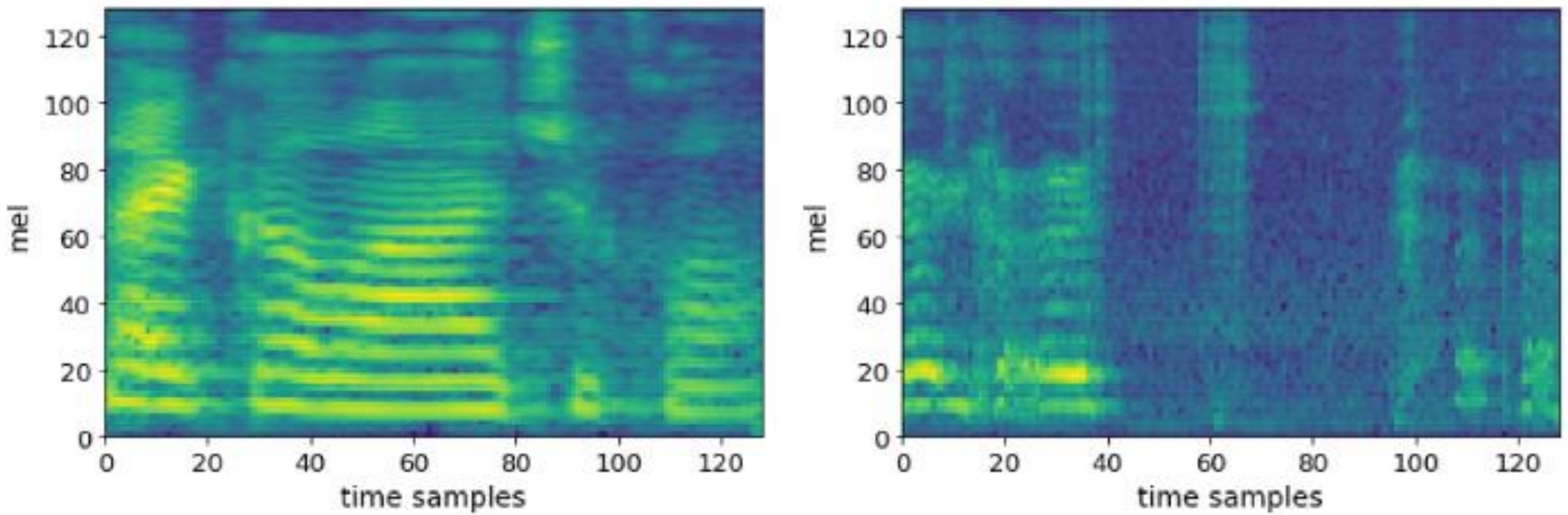
Proposed Model Block Diagram



Attention

- Characteristic embeddings are in **short segments of the whole** utterance
- An expert with knowledge of **where to look** and **what to look for** would be able to perceive these embeddings

- **Hypothesis:** the attention mechanism could locate the salience periods from the spectrograms and could leverage the dysarthria severity recognition task
- **Possible characteristics**: monotonicity, effects of slurring, long pauses

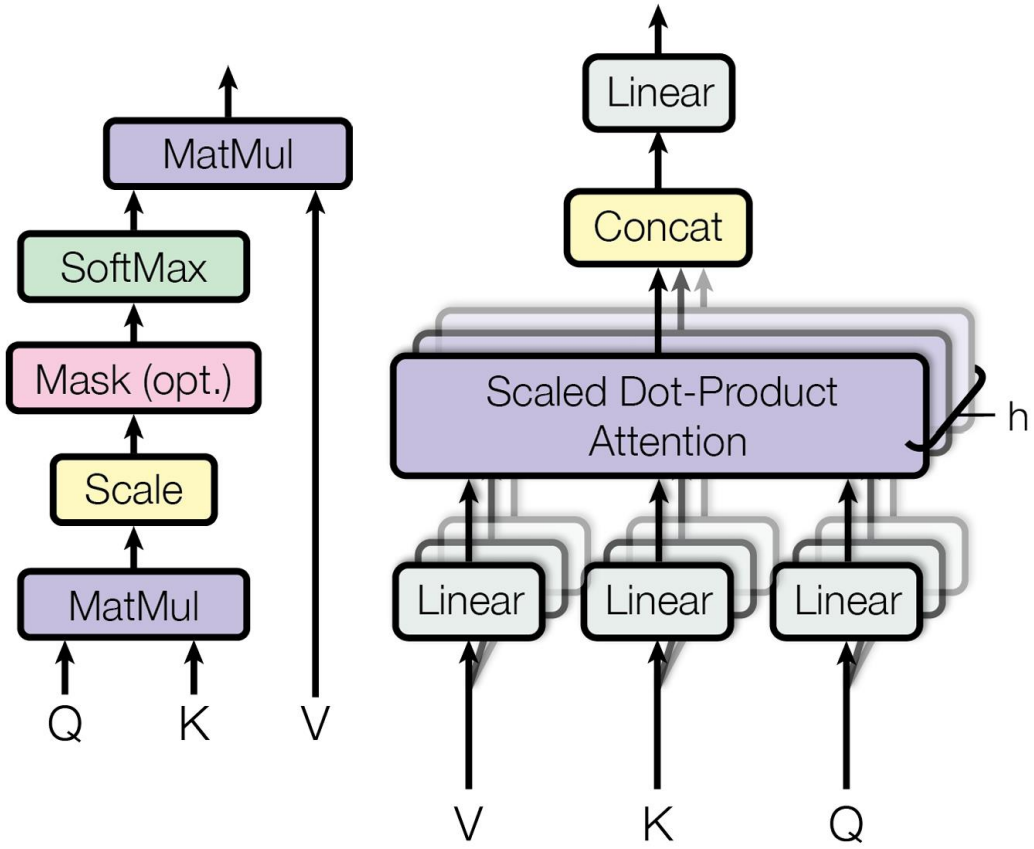


Low severity (left) vs. High severity (right)

Self-attention and Multi-head attention

- Limits of the encoding-decoding
- Self attention mechanism
- Multi head attention mechanism

- **Interpretation**: find the important portions of an image to look at and interpret



Scaled Dot-Product Attention (left) and Multi-Head Attention (right) *Attention Is All You Need, Vaswani et. al (2017)*

Multi-task learning

- Inspired by human learning:
  - Knowledge from **different correlated tasks** is integrated
- Improves data efficiency
- Can lead to faster learning for related tasks under data stringent conditions

- **Hypothesis**: 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

Results: preview

- **95.75%** accuracy of proposed model vs **84.24%** accuracy of baseline model CQT-CNN and **87.14%** accuracy of RES-CNN (ablation)

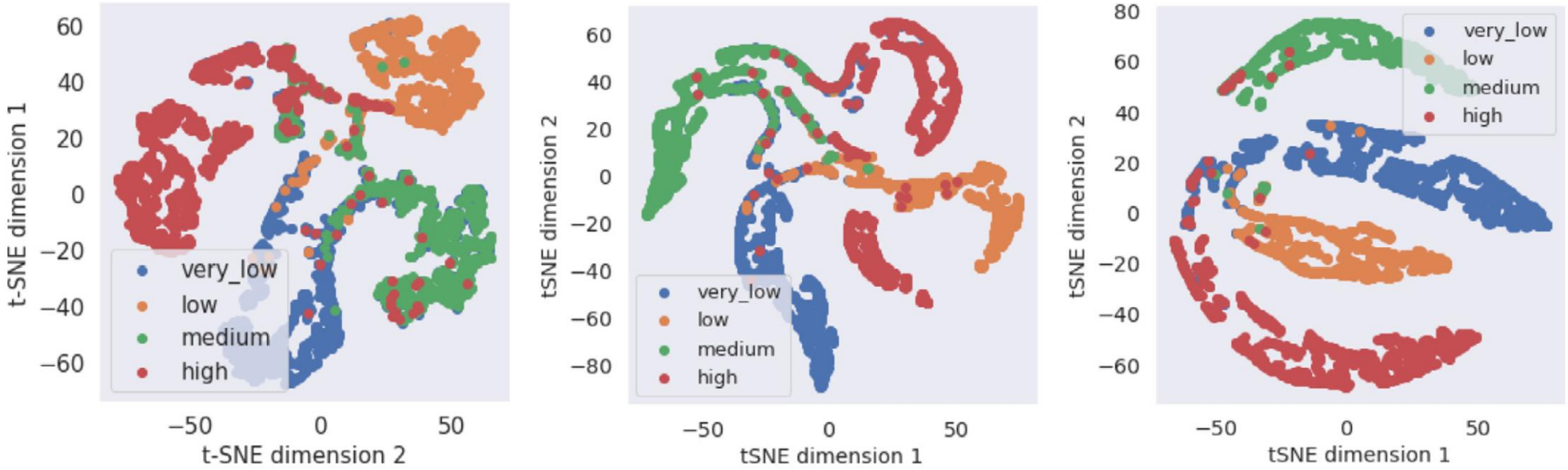
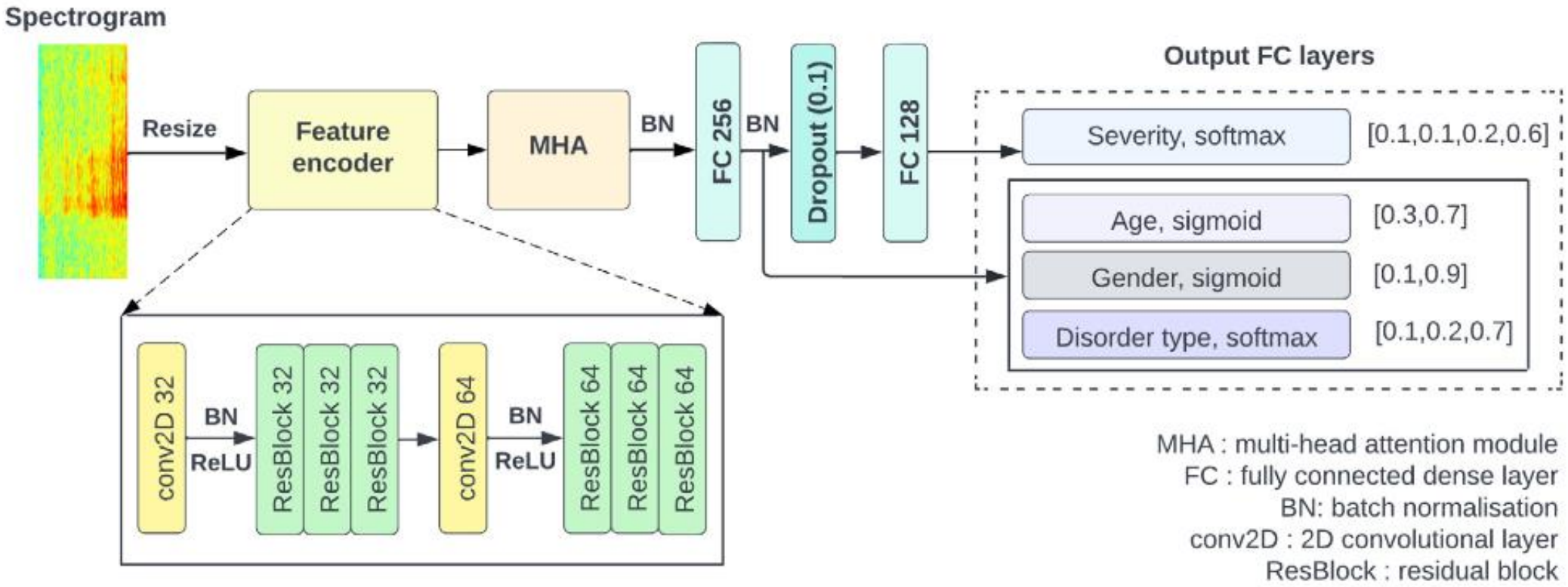
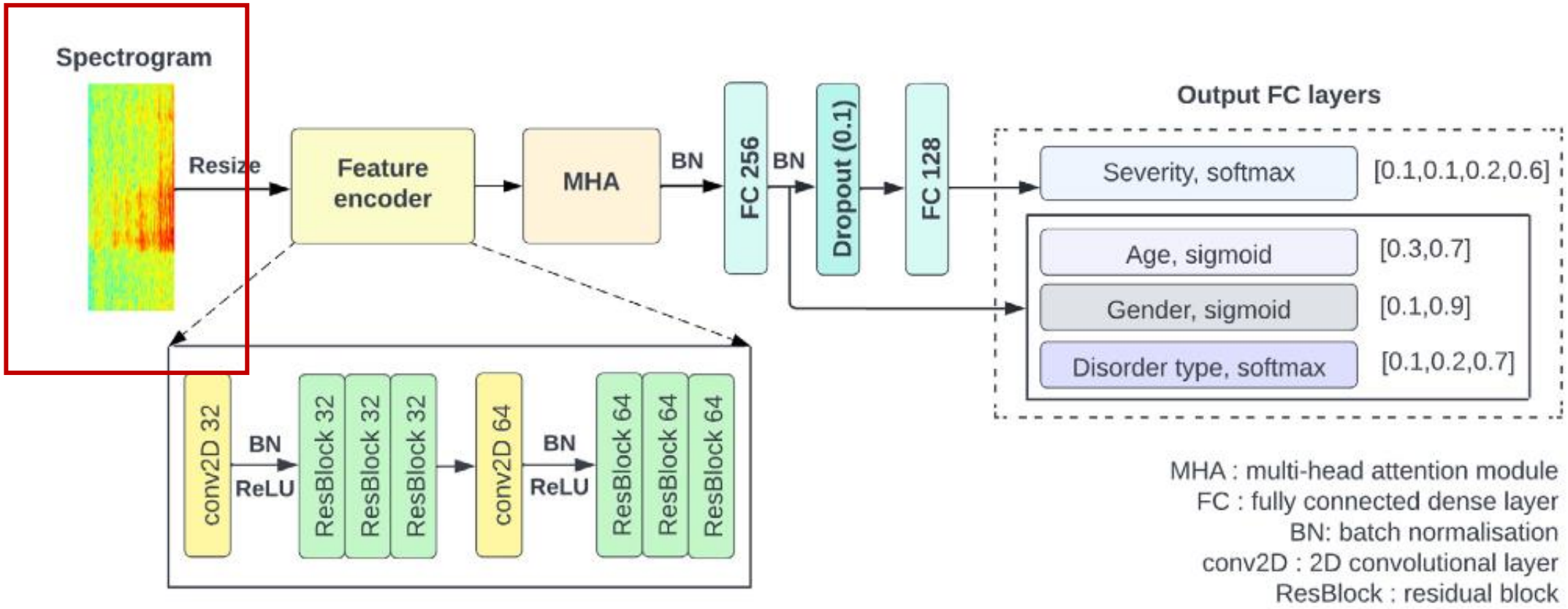


Fig. 4. t-SNE plots of the baseline model (left), ResCNN model (middle) and the proposed model (right).

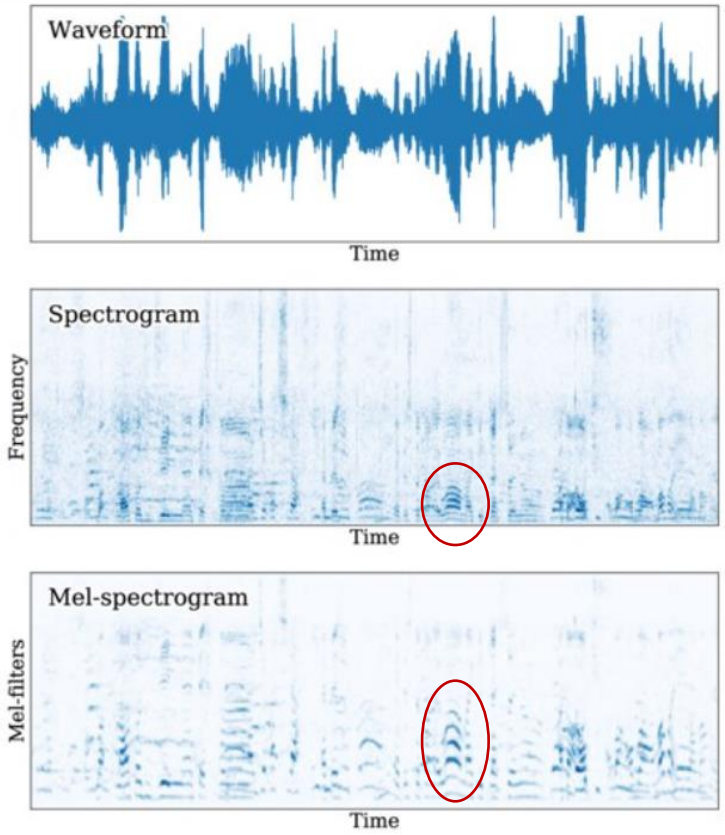
The Proposed Model



Input features



Log mel spectrogram



Dysarthric speech tends to have lower frequency content

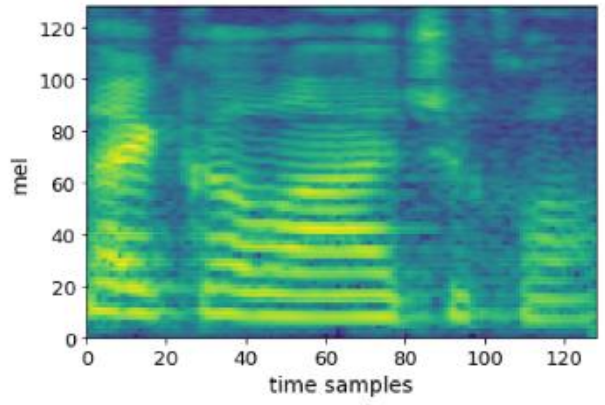
- Mel scale has higher resolution in the lower frequencies, like human hearing
- **Intuitively**: intelligibility rate is a perception task

Exploring convolutional, recurrent, and hybrid deep neural networks for speech and music detection in a large audio dataset, de Benito et. al

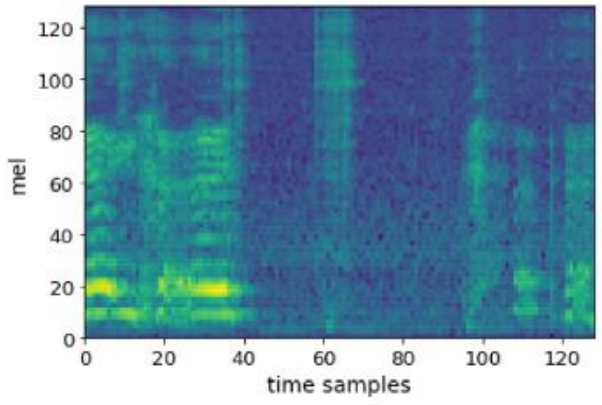
Resized spectrogram

- Silence trimmed instead of clipping to constant length!
- Resized to 64x64 dimension for the CNN classifier

- Poor articulation is visible in the reduced sharpness of the spectrogram

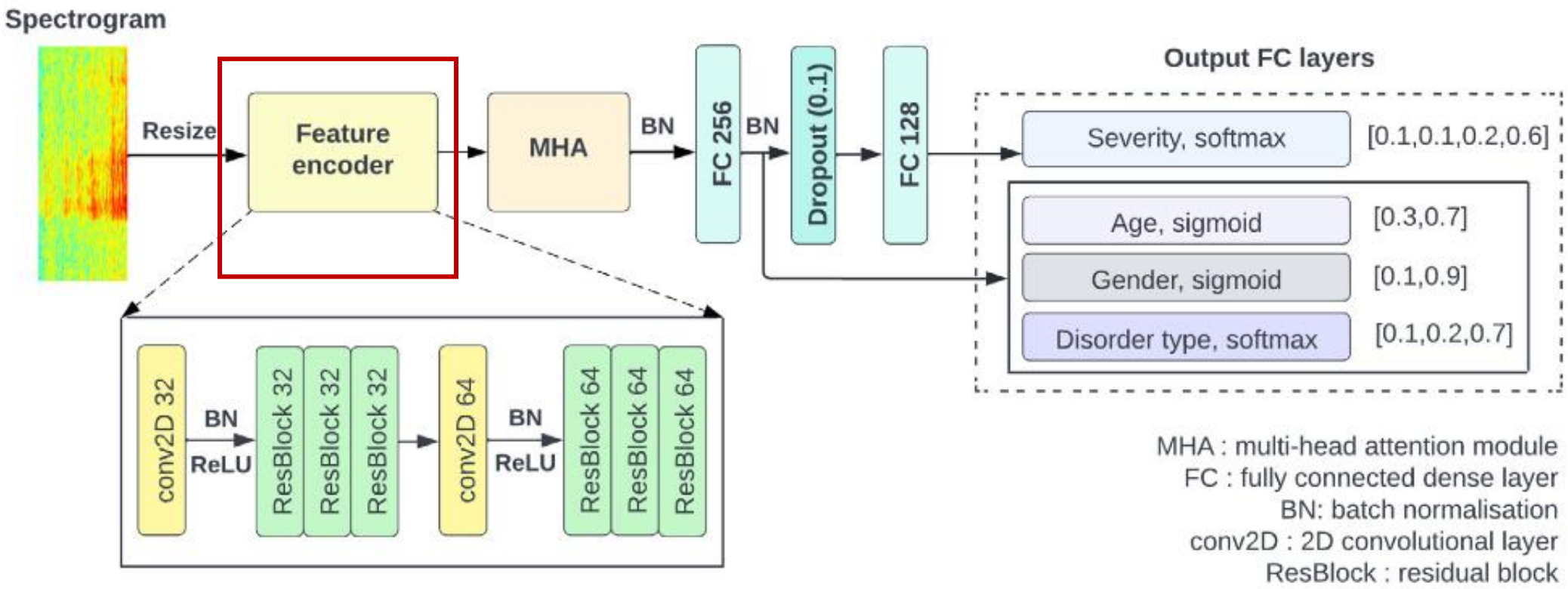


Low severity



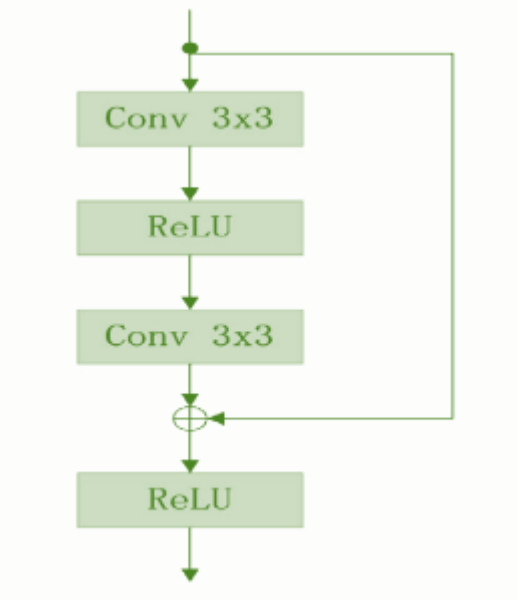
High severity

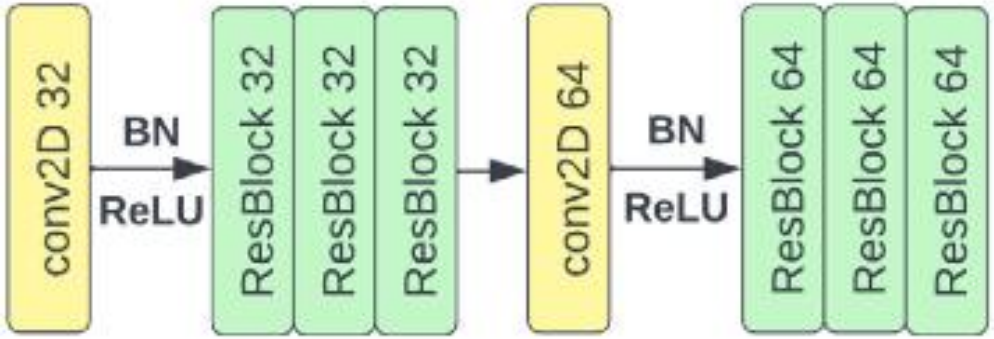
Feature encoding



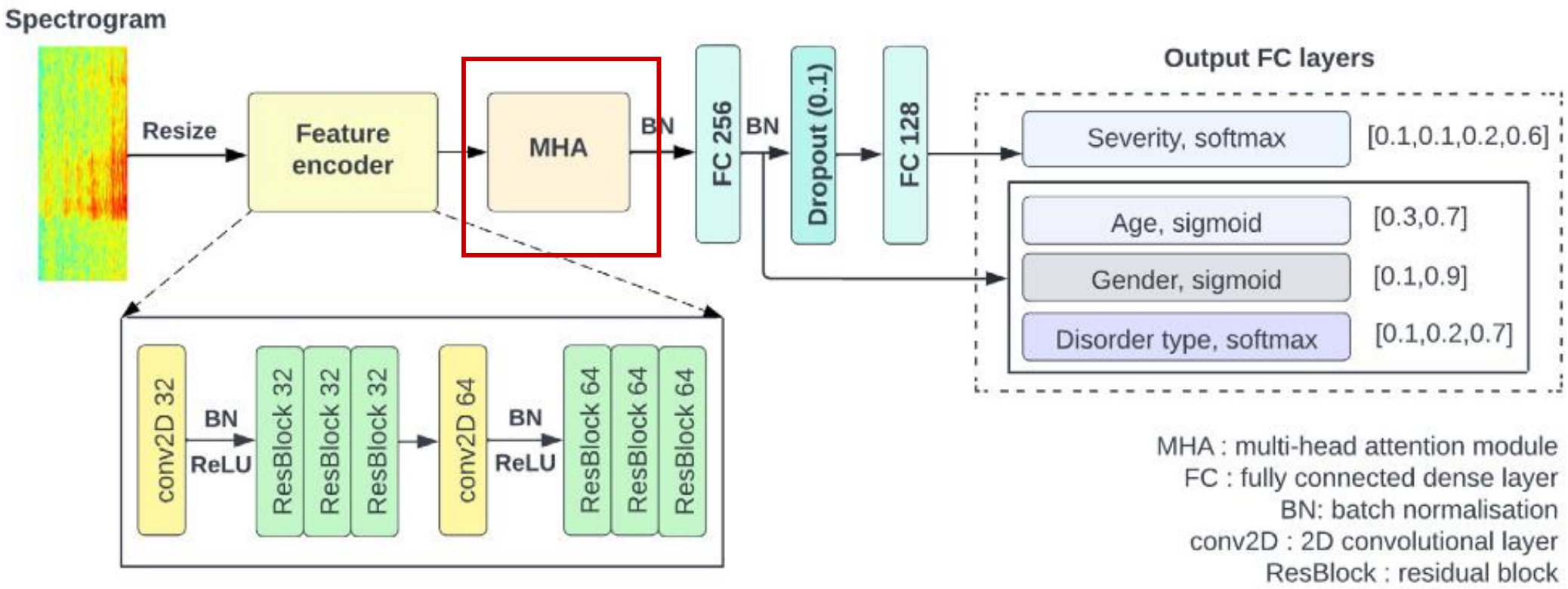
CNN encoder

- **Motivation**: extraction of salient features from the resized spectrogram
- Convolutional filters (32, 64): size 5x5, stride 2x2
- Adjusted version of the Deep Speaker
  - ResCNN-based network



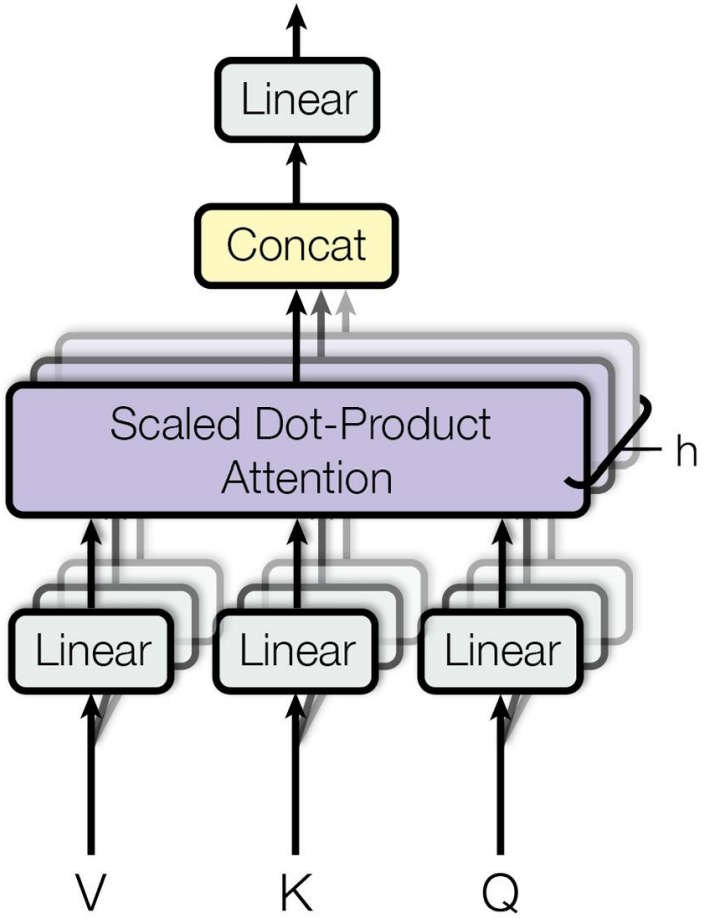


Multi-head attention module



Multi-head attention module

- Three attributes: **Q**ueries, **V**alues and **K**eys
- Multiple projections (**h**)
  - Different projections from the same input
  - Parallel computing
- Allows for recognition of varied dysarthric speech charectaristics present in different severity levels



Mathematical description of MHA

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

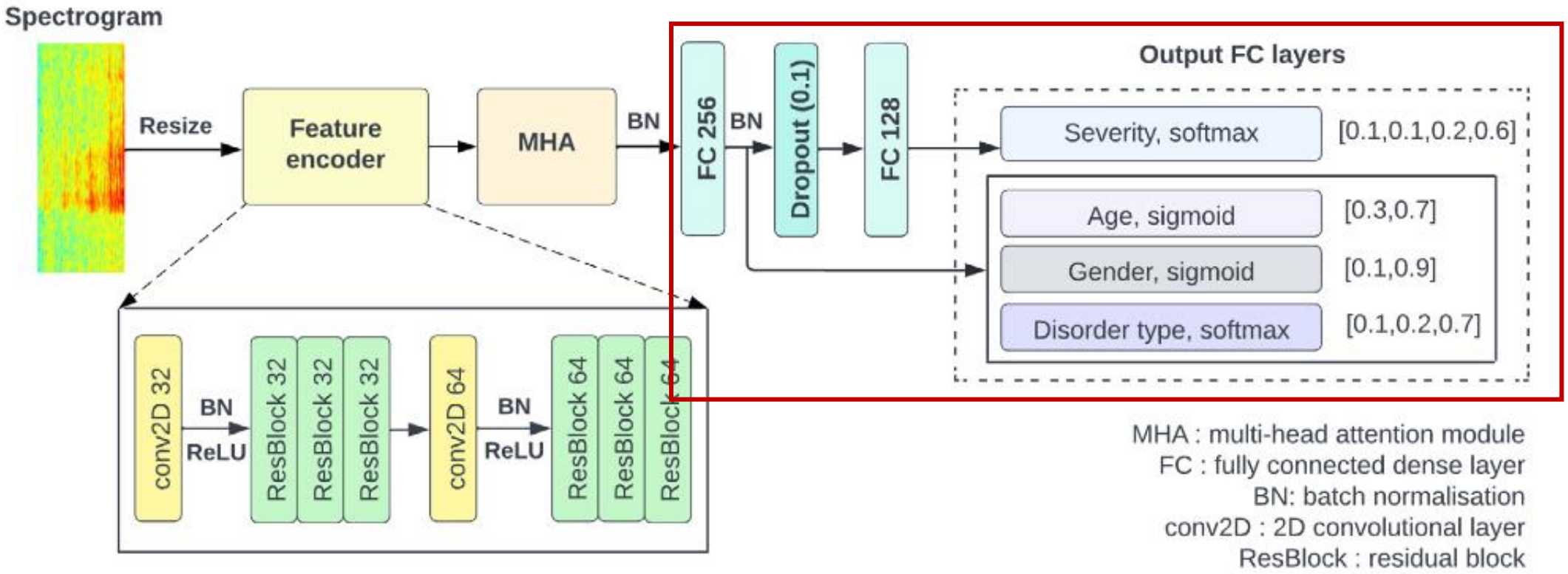
$$MHA(Q, K, V) = concat(head_1, \dots, head_h)W^O$$

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

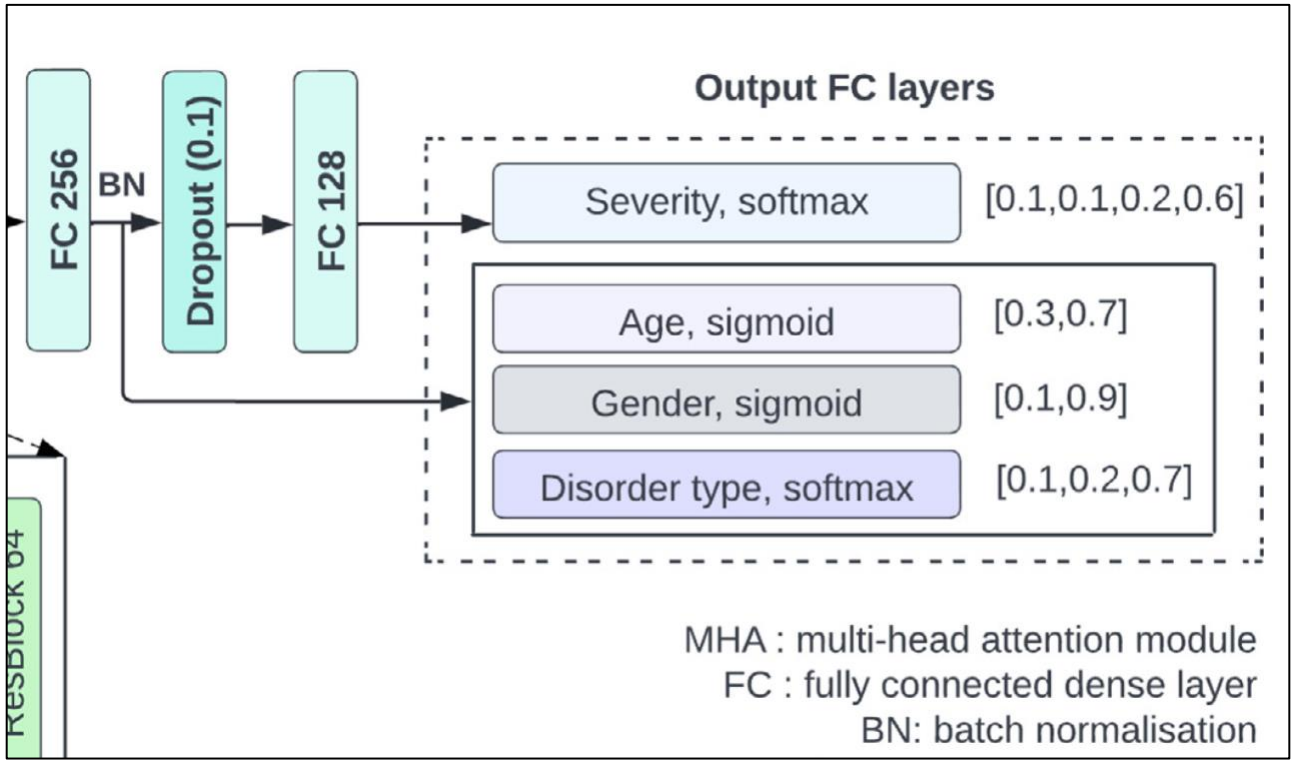
Q – queries, ***K*** – keys, ***V*** – values, ***W*** – projection matrix

d – dimensions, ***h*** – projections

Classifier Neural Network



Description of MTL loss



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

Experiment

- UA-Speech Database; no healthy speakers!
- Training with common words, testing with uncommon words
  - Robustness measure
  - Per speaker: 465 words for training, 300 words for testing
- Stochastic Gradient Descent with momentum 0.9 for 60 epochs

- **Proposed Model**
  - Short-Time Fourier Transform on the log mel scale
  - CNN only for feature extraction

- **Baseline model**: CQT-CNN
  - Constant-Q transform
  - CNN with two hidden layers

Experiment: the speakers of UA-Speech

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

Results: hyper-parameters

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**

- Impact of number of attention heads
- Impact of loss weights
  - Hard parameter sharing prevents overfitting
  - Tuning loss weights

Results: ablation study

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**

- More tasks leads to less overfitting
- Negative transfer of auxiliary learning tasks

Results: confusion matrices

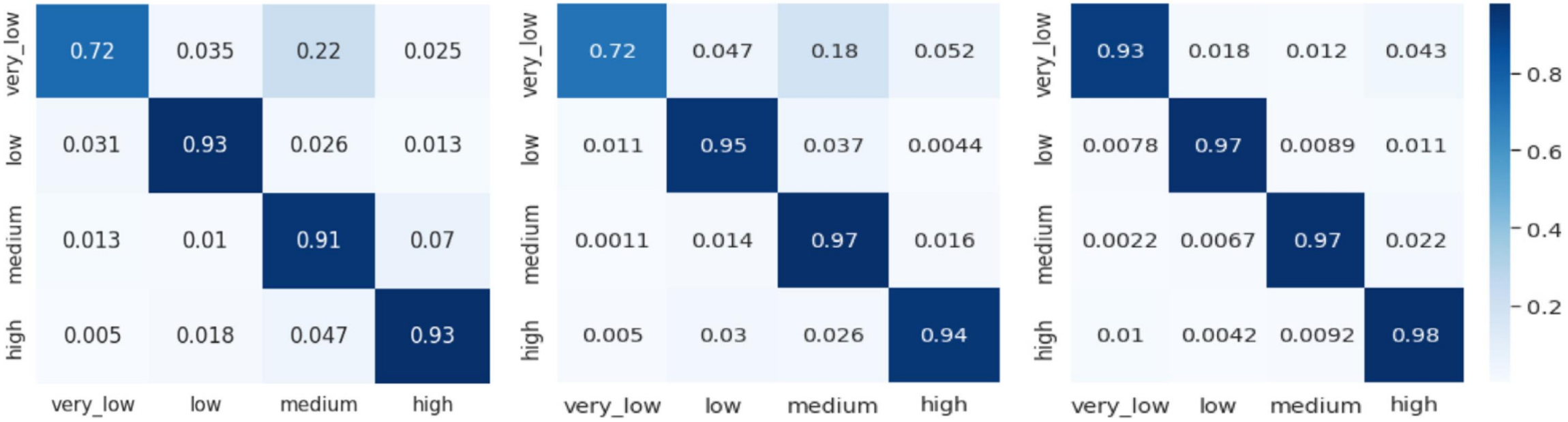


Fig. 5. Confusion matrix given by the baseline model (left), ResCNN model (middle) and the proposed model (right).

- Major recall gain in the ‘very low’ severity class

Results: t-SNE clustering

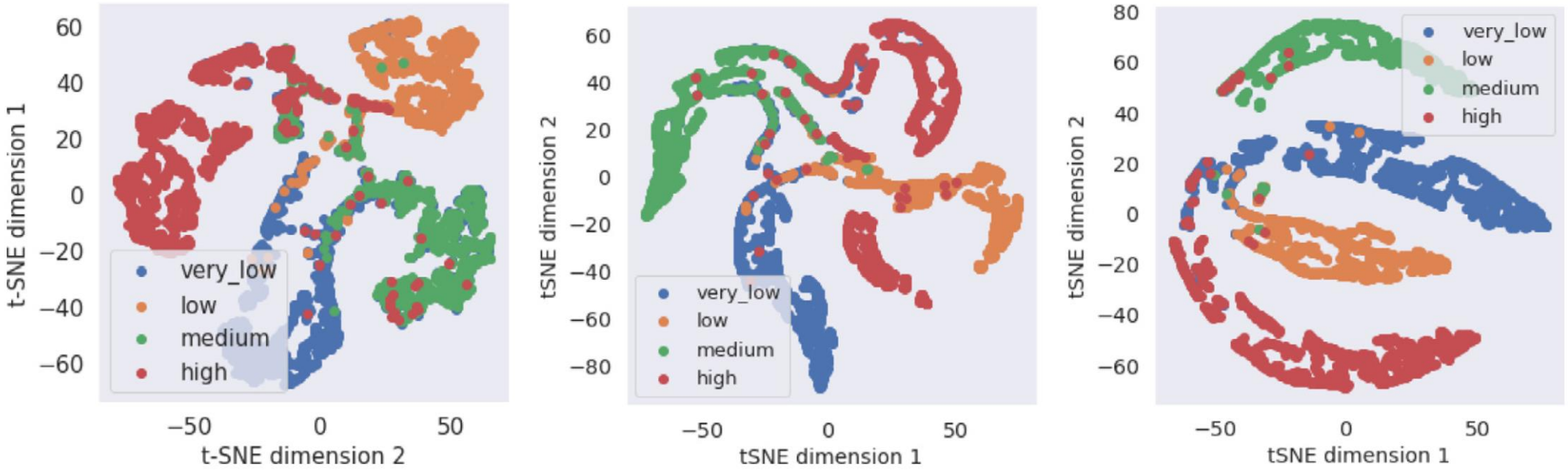


Fig. 4. t-SNE plots of the baseline model (left), ResCNN model (middle) and the proposed model (right).

- Better differentiation for severity classes

Results: contingency tables

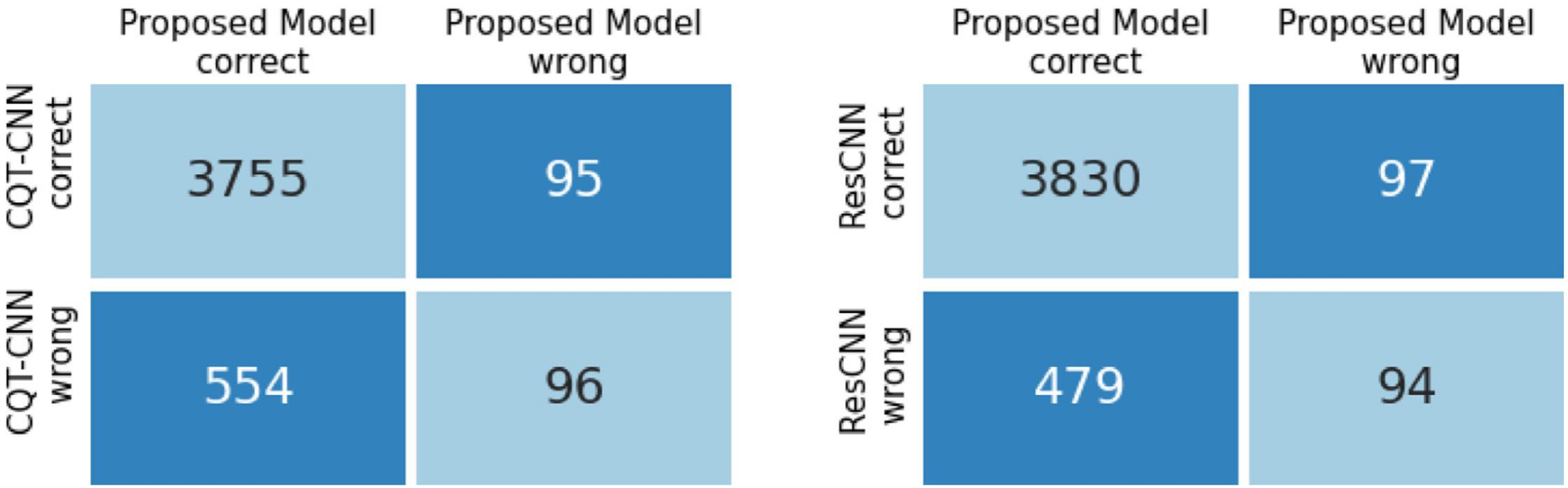


Fig. 6. Contingency tables given by the proposed model against the CQT-CNN model (left) and the ResCNN model (right).

- Results are statistically significant (α = 0.05) by the t statistic

Results: statistical measures

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

Experiment 2: Speaker-dependency check

- Speaker Independent (SID) setting
- Leave One Speaker Out (LOSO)
- Acceptable results for border classes, poor results for intermediate classes; UA-Speech is unbalanced
- Test 1: known words, Test 2: for unknown words

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**

Discussion

- Tripathi et. al 2020b report an accuracy of 54% in the SID setting
  - 1000 hours of training data vs 17 hours
- MHA and MTL can improve performance in **data-scarce** situations
- Age was found not to correlate with severity

Conclusion and Future Work

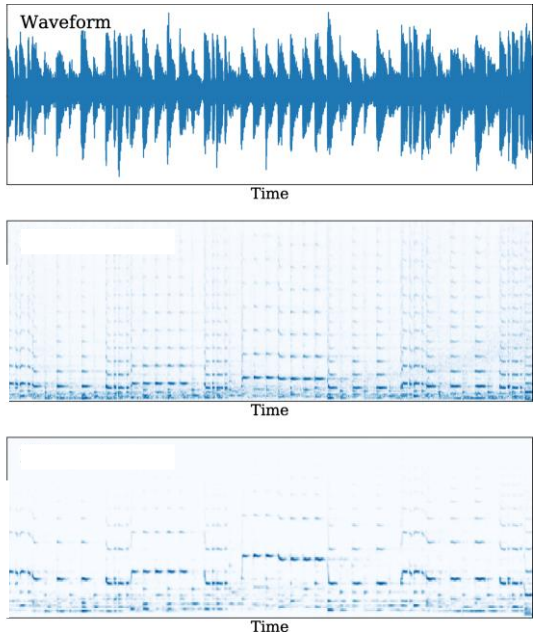
- MHA and MTL: or the **joint learning of different subspace representations** is novel and promising approach in enhancing the performance of dysarthria severity classification in **data-scarce** settings.

- Better time-frequency representation
  - Gabor spectrograms
- Residual networks with squeeze-excitation
- Data augmentation

Thank you for listening!

Assignment

- 1. Given the figures below, which of them in a spectrogram on the normal scale and which is the spectrogram on the mel scale? Why is the mel scale important?



- 2. What is the main limitation of the UA-speech database?
- 3. How does the paper address this limitation?