29th EUSIPCO

eusipco2021.org

StutterNet: Stuttering Detection Using Time Delay Neural Network

Shakeel A. Sheikh[1], Md Sahidullah[1], Fabrice Hirsch[2], Slim Ouni[1]

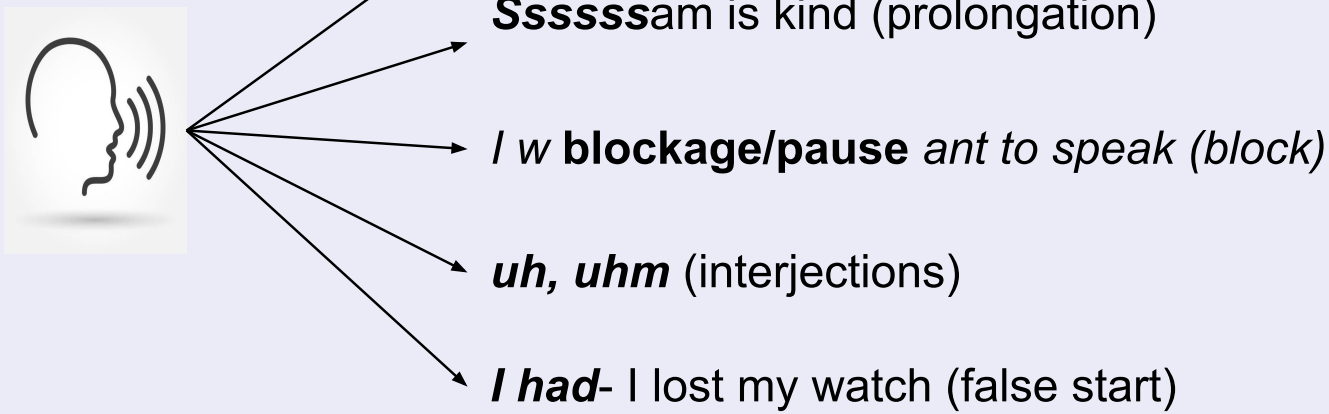
Université de Lorraine, CNRS, Inria, LORIA, F-54000, Nancy, France[1]
Université Paul-Valéry Montpellier, CNRS, Praxiling, Montpellier, France[2]

23-27 August, 2021

Stuttering problem

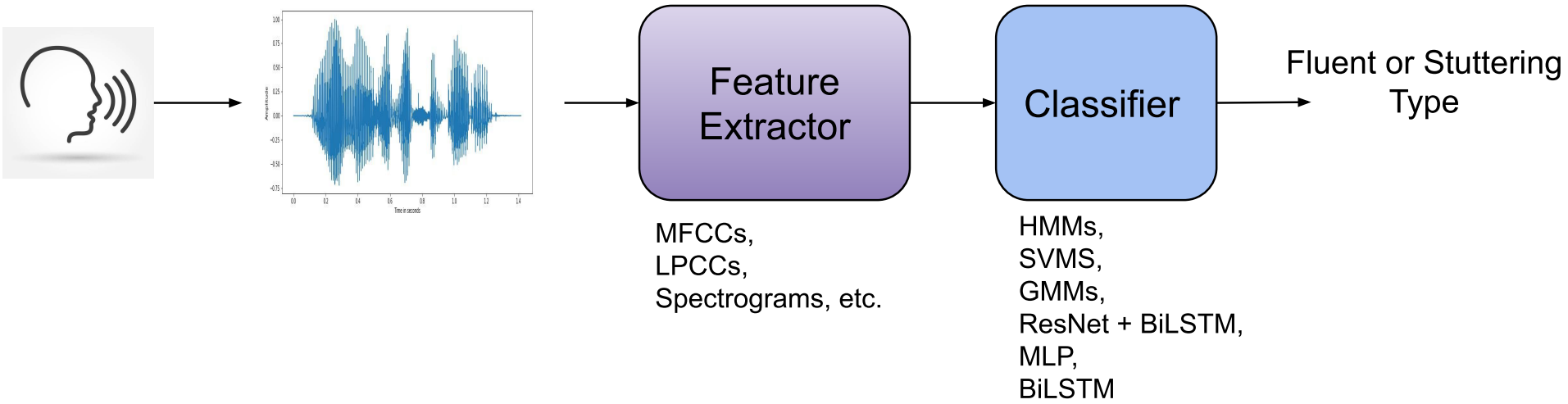
Definition

- A speech disorder, also known as *stammering*.
- Characterized by repetition of sounds, prolongation of sounds; and interruptions in speech known as blocks.



1. Barry Guitar. Stuttering: An integrated approach to its nature and treatment. Lippincott Williams & Wilkins, 2013.

Stuttering detection system



2. T. Kourkounakis et al, "Detecting multiple speech disfluencies using a deep residual network with bidirectional long short-term memory". In: ICASSP 2020, pp. 6089–6093.
3. C. Lea, et al, "SEP-28k: A dataset for stuttering event detection from podcasts with people who stutter," ICASSP 2021, pp. 6798-6802.

Time delay neural network (TDNN)

- TDNN is a neural network architecture used for modeling temporal data with contextual information.
- Applications:
  1. speech recognition,
  2. emotion detection,
  3. speaker recognition, etc.

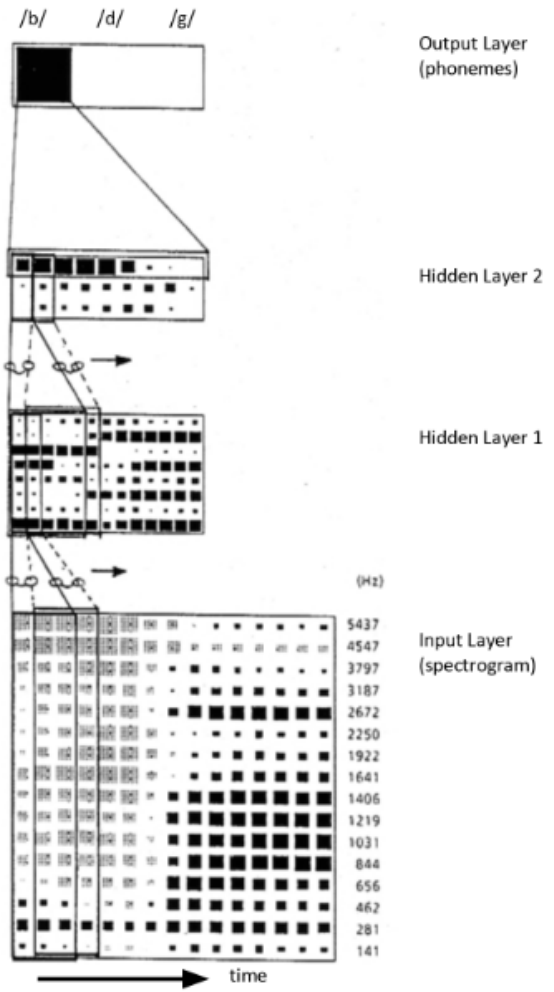
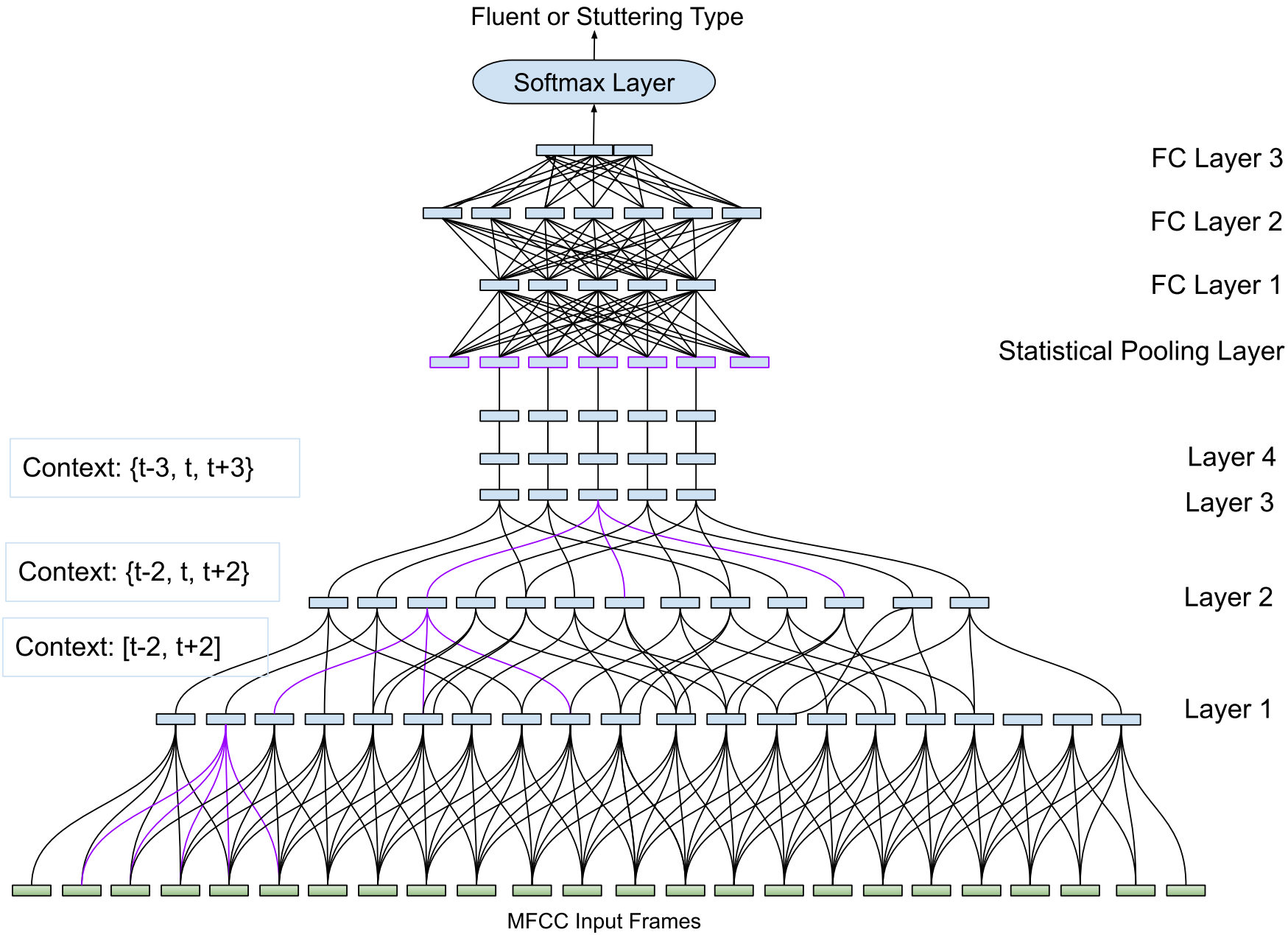


Fig. taken from [4].

4. Alex Waibel et al. "Phoneme recognition using time-delay neural networks". In: IEEE Transactions on Acoustics, Speech, and Signal Processing 37.3 (1989), pp. 328–339.

StutterNet architecture



StutterNet layer configuration

StutterNet with input: D-dim (MFCC) X *T* (Chunk size)

Layer	Input X Output	Context
TDNN1	D X 512	[t-2, t+2]
TDNN2	1536 X 512	{t-2, t, t+2}
TDNN3	1536 X 512	{t-3, t, t+3}
TDNN4	512 X 512	{t}
TDNN5	512 X 1500	{t}
Statistical Pooling	1500*T* X 3000	[0, *T*)
FC1	3000 X 512	-
FC2	512 X 512	-
FC3	512 X *N*	-

- FC: Fully connected layer.
- TDNN and FC are followed by a ReLU activation function and batch normalization.
- N is number of classes.

Configuration

Features

- Audio clip 4 sec.
- 20-dim MFCC input features, with Hamming window of 20 ms and hop-length of 10 ms.

StutterNet parameters

- Learning rate $= 10^{-4}$.
- Optimizer $=$ Amsgrad.
- Batch size $= 64$.
- Cross entropy loss function.
- Early stopping criteria with patience seven on validation loss.

Dataset and experimental setup

- *UCLASS*, 128 speakers, 110 Males, 18 Females, 4674 samples.
- #classes = four (blocks, prolongations, repetitions and fluents).
- Train Speakers: 80%, Val Speakers: 10%, Test Speakers: 10%.
- K-fold cross validation with K=10.
- Evaluation metrics: Accuracy (Acc), Mathews correlation coefficient (MCC), precision, recall, F1-score.

Results

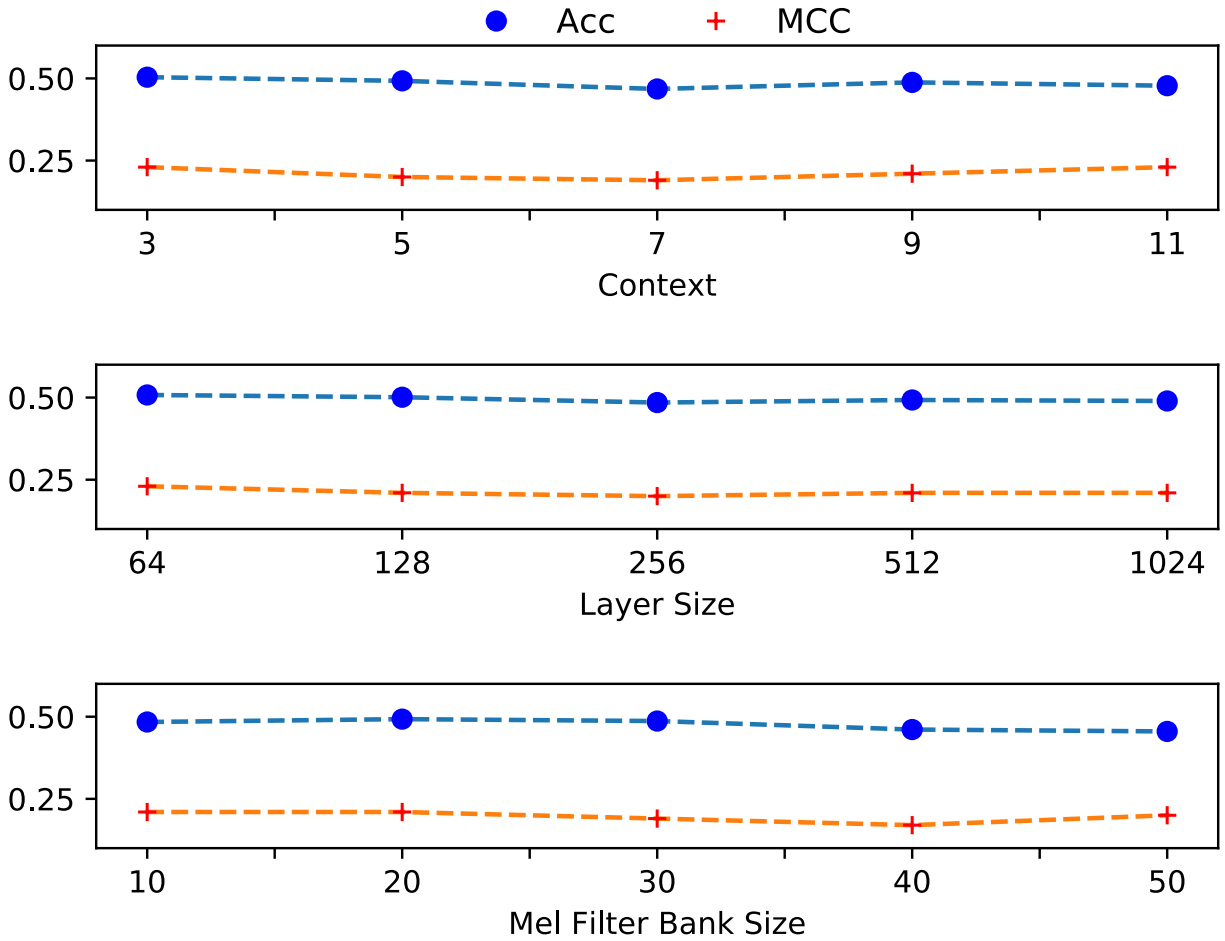
	Precision			
Method	Repetition	Prolongation	Block	Fluent
ResNet+BiLSTM[5]	0.33	0.42	0.43	**0.63**
StutterNet (Baseline)	**0.36**	**0.43**	0.42	0.59

	Recall			
Method	Repetition	Prolongation	Block	Fluent
ResNet+BiLSTM[5]	0.20	**0.23**	**0.53**	0.55
StutterNet (Baseline)	**0.28**	0.17	0.42	**0.67**

	F1-Score			
Method	Repetition	Prolongation	Block	Fluent
ResNet+BiLSTM[5]	0.22	**0.28**	0.44	0.52
StutterNet (Baseline)	**0.30**	0.23	0.42	**0.62**

5 T. Kourkounakis et al, "Detecting multiple speech disfluencies using a deep residual network with bidirectional long short-term memory". In: ICASSP 2020. 2020, pp. 6089–6093[1].

Impact of context, layer size and MFCC size



Optimal configuration: layer size = 64, context = 5, MFCC = 20

StutterNet (optimized)

	Precision			
Method	Repetition	Prolongation	Block	Fluent
ResNet+BiLSTM[5]	0.33	0.42	0.43	**0.63**
StutterNet (Optimized)	0.35	0.31	**0.47**	0.59

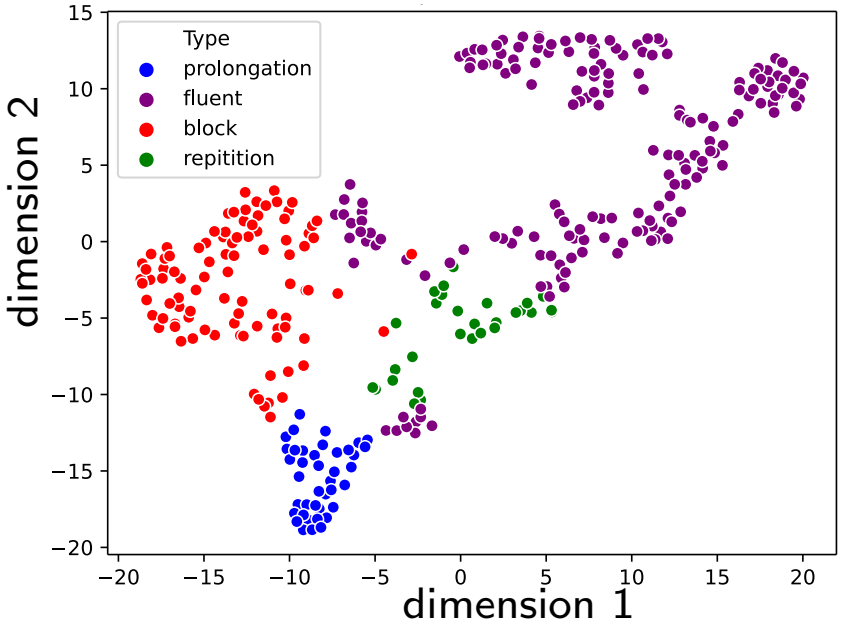
	Recall			
Method	Repetition	Prolongation	Block	Fluent
ResNet+BiLSTM[5]	0.20	**0.23**	**0.53**	0.55
StutterNet (Optimized)	0.24	0.13	0.47	**0.70**

	F1-Score			
Method	Repetition	Prolongation	Block	Fluent
ResNet+BiLSTM[5]	0.22	**0.28**	0.44	0.52
StutterNet (optimized)	0.27	0.16	**0.46**	**0.63**

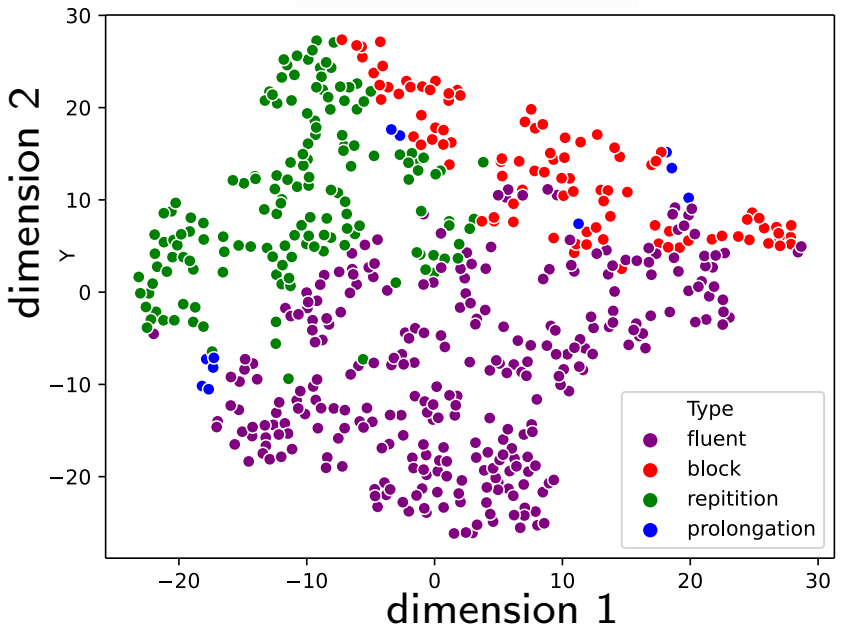
Method	Accuracy				Tot. Acc.	MCC.
	Rept	Pr	B	F		
Resnet+BiLSTM[5]	20.39	**23.17**	**53.33**	55.00	46.10	0.20
StutterNet (Optimized)	**23.98**	12.96	47.14	**69.69**	**50.79**	**0.23**

t-SNE plot

ResNet+BiLSTM



StutterNet



Conclusion

- Single unified and ASR free system to detect all stuttering types and fluent segments.
- Relative improvement in Acc = 10.17% and MCC = 15%
- Significantly lesser #parameters to train, 71K(*StutterNet*) in comparison to 24M (ResNet+BiLSTM).
- Generalization of *StutterNet* for cross corpora scenario with variable accents.
- Precise identification of stuttering event can't be predicted.
- Loss of information like pitch, phase, in extracting MFCC features.
- End-to-end systems for stuttering detection, where network may learn stuttering related features from raw signal.

Appendix: Evaluation metrics

MCC

$$MCC = \frac{cs - \mathbf{t}.\mathbf{p}}{\sqrt{s^2 - \mathbf{p}.\mathbf{p}}\sqrt{s^2 - \mathbf{t}.\mathbf{t}}} \tag{1}$$

where,

- $s = \sum_i \sum_j C_{ij}$, is the total number of samples,
- $c = \sum_k C_{kk}$, is the total number of samples correctly predicted,
- $p_k = \sum_i C_{ki}$, is the number of times class k was predicted,
- $t_k = \sum_i C_{ik}$, is the number of times class k truly occurred.

6. Gorodkin, J. (2004). Comparing two K-category assignments by a K-category correlation coefficient. Computational Biology and Chemistry, 28(5-6), 367-374.