Stuttering Identification using Deep Learning

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

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

Abstract

Stuttering identification (SI) is a speech characterization problem that has been approached via different signal processing and statistical machine learning methods. Speech technology has been drastically revolutionized, thanks to advances in deep learning but SI has received less attention. This work explores different deep learning algorithms to solve the SI problem. First, we introduce StutterNet, a time-delay neural network architecture for SI. Then, we investigate multi-task (MTL) and adversarial (ADV) learning frameworks to learn robust speech representation. To address the limited data problem, we further introduce speech embeddings for SI where embeddings were extracted from models trained on large datasets and for separate tasks. We have achieved the best SI performance so far using the Wav2Vec2.0 embeddings with the neural network backend.

Introduction

- Stuttering is a neuro-developmental speech impairment defned by blocks, prolongation, repetitions and interjections.
- Approximately 70 million ≈ 1% World's population suffer from stuttering.
- The stuttering detection system commonly consists of an acoustic feature extractor followed by a statistical classifier.

Motivation

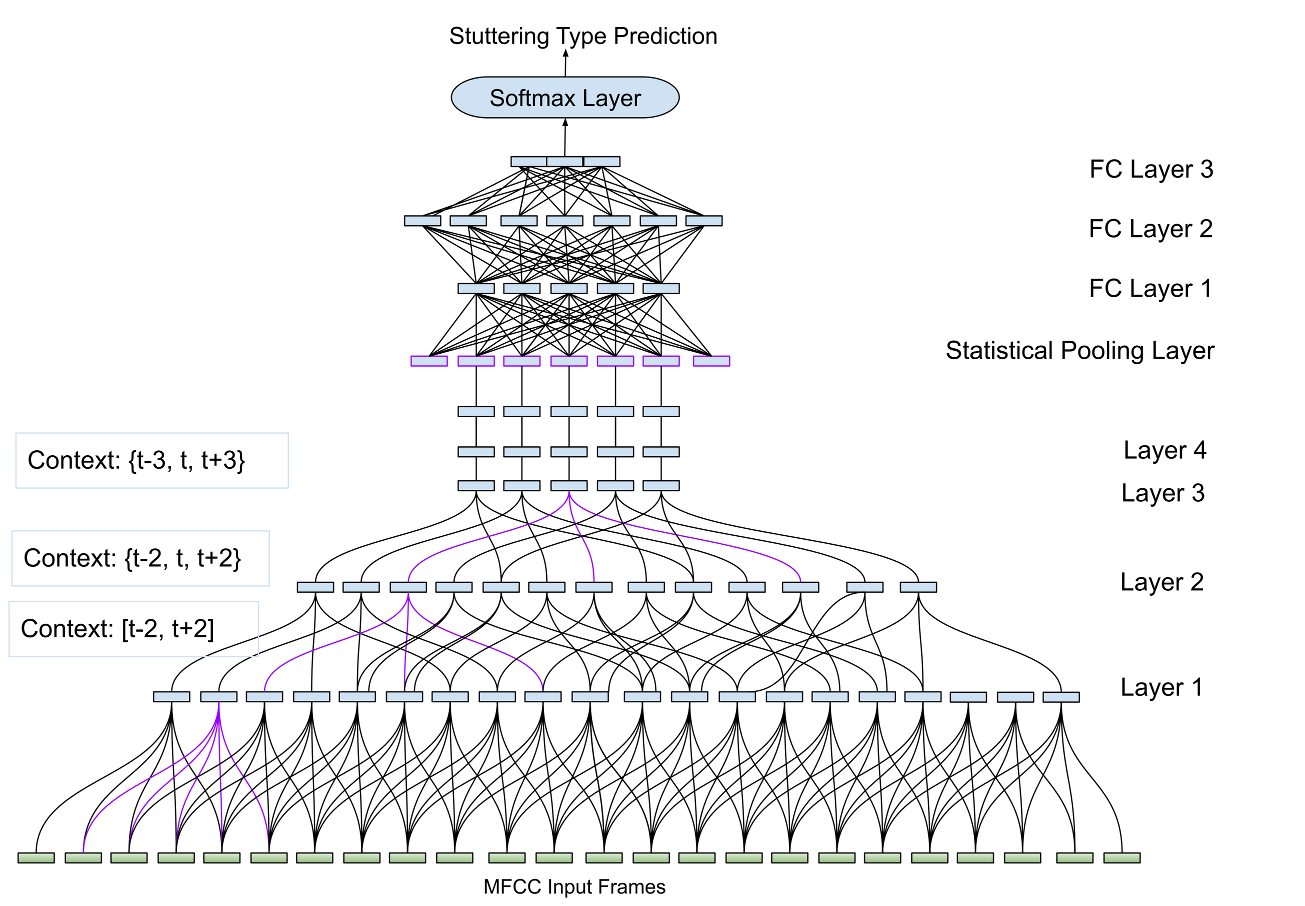
- Nearly impossible to access virtual assistants like Alexa, Apple Siri.
- Helpful for speech therapists.
- The success of statistical machine learning methods in limited in SI due to its complex nature among the different disfluent categories.
- Deep learning has been successfully applied in different speech classification tasks such as automatic speech recognition and speaker recognition but has not been much explored in SI.

Main Contributions

1. *StutterNet*: an end-to-end architecture for SI.
2. *Advancing StutterNet via multi-tasking (MTL) and adversarial (ADV) learning for robust SI*
3. *Stuttering identification with speech embeddings.*

Neural Network Architecture

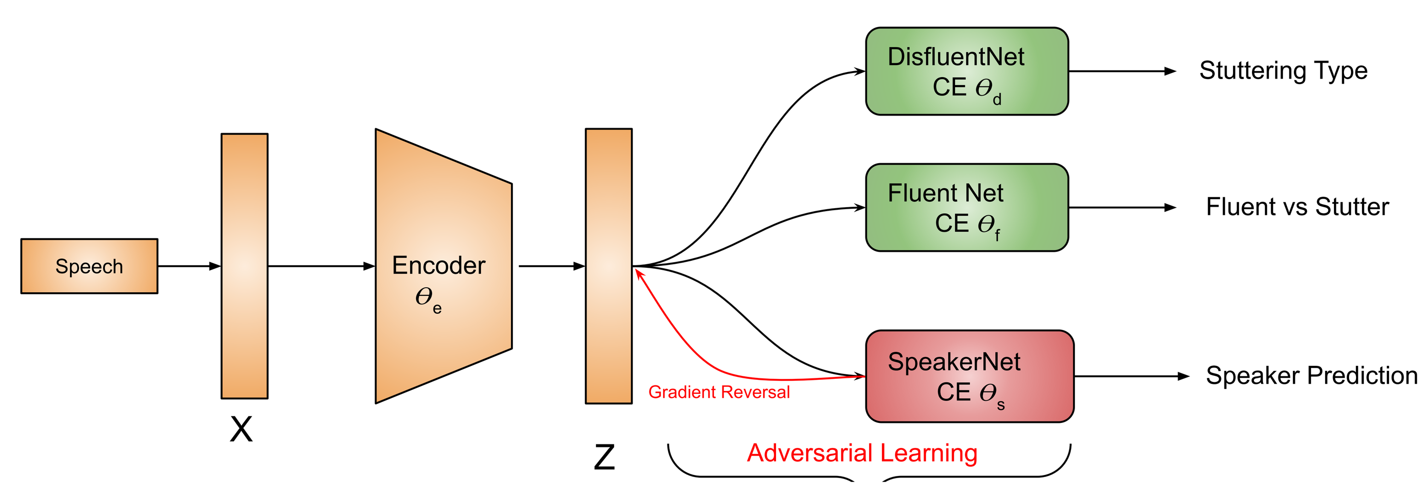
StutterNet:

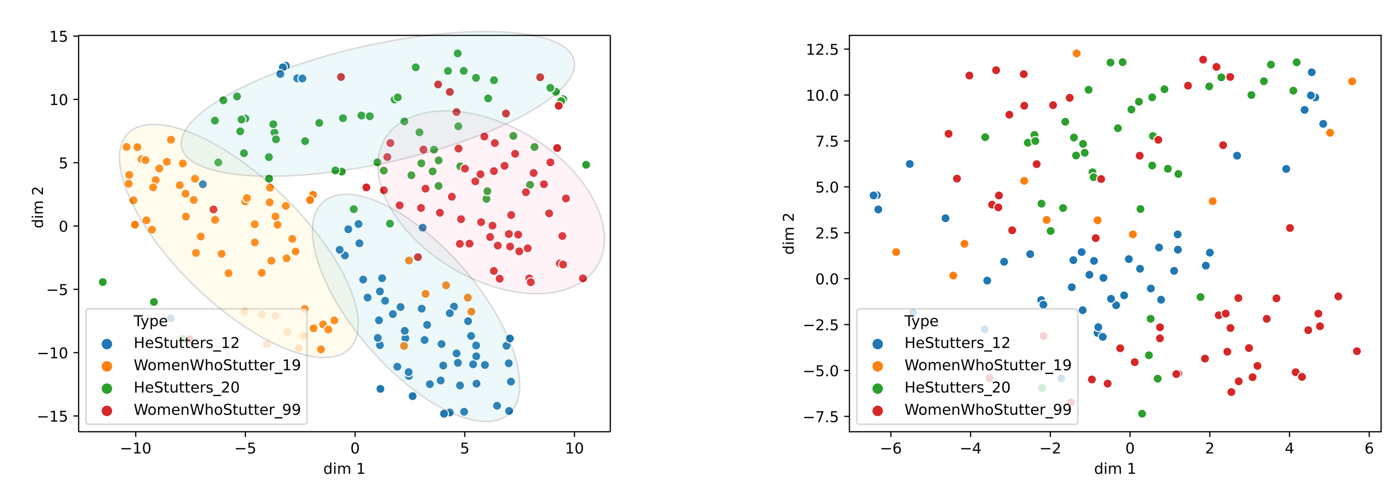


Multi-task and Adversarial Learning:

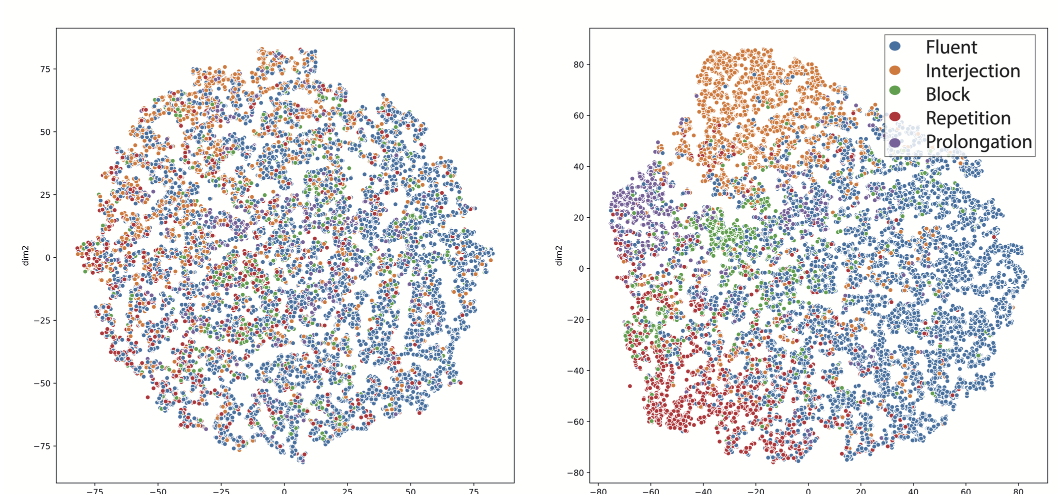
MTL objective function

$$\mathcal{L}(\theta_e, \theta_f, \theta_d, \theta_s) = (1-\lambda) * \mathcal{L}_{\text{stutter}}(\theta_e, \theta_f, \theta_d) + \lambda * \mathcal{L}_{\text{speaker}}(\theta_e, \theta_s)$$
$$\mathcal{L}_{\text{stutter}}(\theta_e, \theta_f, \theta_d) = \mathcal{L}_{\text{fluent}}(\theta_e, \theta_f) + \mathcal{L}_{\text{disfluent}}(\theta_e, \theta_d)$$



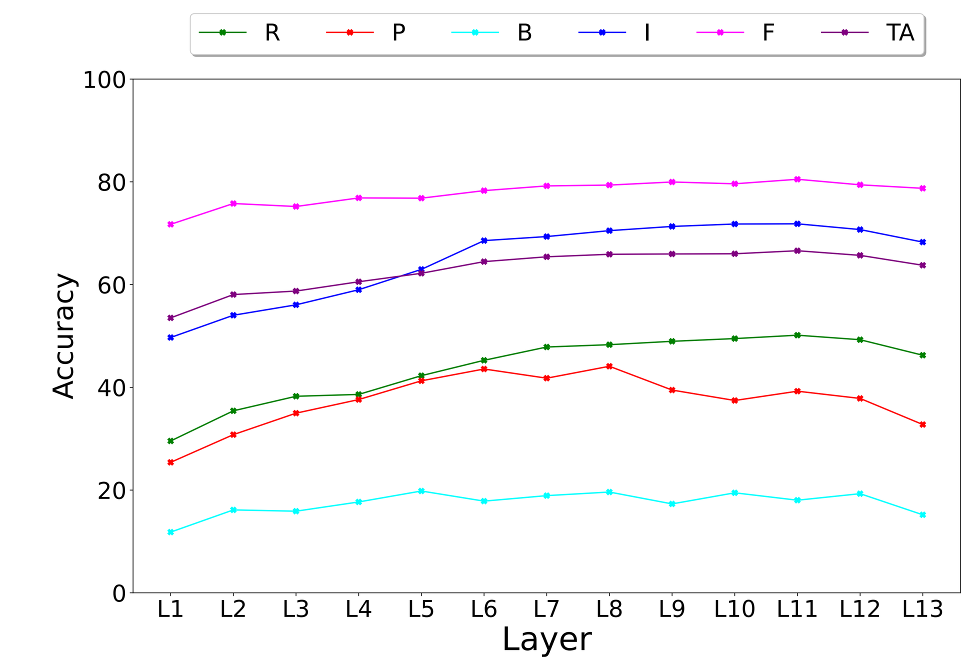


Speech Embeddings



ECAPA-TDNN Wav2Vec2.0

- SOTA speech embeddings like ECAPA-TDNN and Wav2Vec2.0 are widely used in speech tasks.
- Analyzed Wav2Vec2.0 speech embeddings are more suitable for SI.



Results and Discussion

Table 1. SD results on SEP-28k dataset (TA: Total accuracy, B: Block , F: Fluent , R: Repetition , P: Prolongation , I: Interjection, BL: Baseline, NN: 3 layered fully connected neural network).

Model	R	P	B	I	F	TA
StutterNet [3]	21.99	27.78	1.98	49.99	88.18	60.33
BL (Multi Branch)	28.70	37.89	9.58	57.65	74.43	57.04
MB StutterNet + MTL	31.59	31.62	10.23	58.92	72.14	56.09
MB StutterNet + ADV	27.24	32.89	8.33	56.36	77.10	57.51
NN + Wav2Vec2.0	46.79	40.79	**23.86**	69.54	**84.32**	**68.35**

- BL shows a relative improvement of 26% in disfluent classes.
- λ acts a control parameter for the podcast information to flow through the network.
- The well-formed podcast clusters in the MTL indicate that the model is attempting to learn podcast dependent stuttering information. The clusters, on the other hand, are not observable in the adversarial scenario, and the model is attempting to learn these meta-data invariant robust stutter representations.
- Wav2Vec2.0 captures rich stutter discriminative features.
- Overall improvement with Wav2Vec2.0 is 19.83% in SI.

References

[1] Liam Barrett, Junchao Hu, and Peter Howell. Systematic review of machine learning approaches for detecting developmental stuttering. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:1160–1172, 2022.

[2] Shakeel Sheikh, Md Sahidullah, Fabrice Hirsch, and Slim Ouni. Machine learning for stuttering identification: Review, challenges & future directions. *arXiv preprint arXiv:2107.04057*, 2021.

[3] Shakeel Ahmad Sheikh, Md Sahidullah, Fabrice Hirsch, and Slim Ouni. StutterNet: Stuttering detection using time delay neural network. In *Proc. 29th EUSIPCO*, 2021.

[4] Shakeel Ahmad Sheikh, Md Sahidullah, Fabrice Hirsch, and Slim Ouni. Introducing ECAPA-TDNN and Wav2Vec2.0 embeddings to stuttering detection. *arXiv preprint arXiv:2204.01564*, 2022.

[5] Shakeel Ahmad Sheikh, Md Sahidullah, Fabrice Hirsch, and Slim Ouni. Robust stuttering detection via multi-task and adversarial learning. *arXiv preprint arXiv:2204.01735*, 2022.