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 an interdisciplinary research problem in which a variety of research studies (in terms of auditory feature extraction and classification approaches) have been carried out with the goal of creating automated tools for its detection and identification. The majority uses language models to detect and identify stuttering. The speech domain has been drastically revolutionized thanks to advances in deep learning, but SI has received less attention so far. In this work for stuttering identification, we explore time delay neural networks (TDNN) which is suitable for capturing temporal information of disfluent utterances. We also investigate how multi-task (MTL) and adversarial (ADV) learning framework can help to learn robust stuttering features. In addition, we explore different pre-trained speech embeddings in SI and we achieved state-of-the-art performance on a large corpus. The pre-trained speech embeddings based SI methods has shown the best performance with an overall accuracy of 68.35% on SEP-28k dataset.

Introduction

Stuttering is a neuro developmental speech impairment which is caused due to the failure of speech sensorimotors responsible for speech production, and is defined by the disruption of *core behaviours:* prolongations, blocks and repetitions and by uncontrolled filler interjections. Approximately 70 million $\approx$ 1% World's population suffer from stuttering and is found more in males than females with a Male:Female ratio of 4:1.

Motivation

- Automatic recognition systems (ASR) fail to recognize stuttered speech.
- Nearly impossible to access virtual assistants like Alexa, Apple Siri.
- Conventional stuttering identification (SI) is manual, laborious, and time consuming, and is also biased towards the subjective belief of speech therapists.
- Annotation of stuttering samples is ambiguous.
- Exploiting acoustic cues present in the speech utterance.
- Exploiting podcast information via multi-task and adversarial learning in SI.
- Use of transfer learning (Wav2Vec2.0 and ECAPA-TDNN embeddings) in SI due to availability of limited size stuttering datasets.

Main Contributions

1. *StutterNet*: a time delay neural network for SI.
2. *Multi-tasking, and Adversarial Learning for SI*. According to early findings, the MTL improves the detection efficiency of just disfluent classes while increasing the confusion for fluent samples. As illustrated in Fig as well, we observe that the ADV framework learns robust stutter characteristics that are stutter discriminative while also being metada invariant. Moreover, We employ a multi-branch training strategy to address class imbalance in SI.
3. *Stuttering identification with speech embeddings*. We investigate the of role of ECAPA-TDNN speaker and Wav2Vec2.0 speech representations for SI, and to our knowledge, is the first ever study in SI. We also examine the impact of information extracted from various different layers of Wav2Vec2.0. Moreover, we find layer fusion and embedding fusion of information from ECAPA-TDNN and Wav2Vec2.0 further boosts the SI performance.

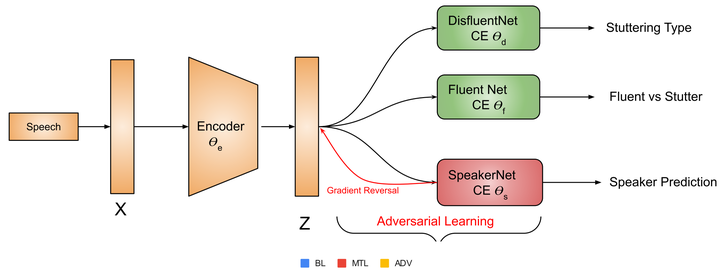
Multi-task and Adversarial Learning

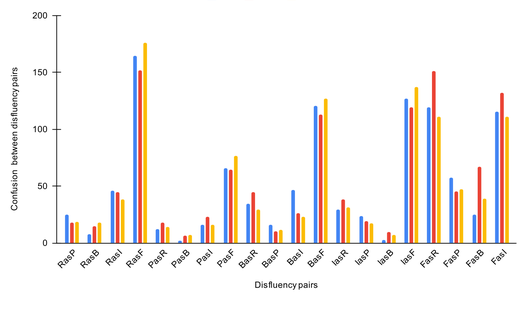
MTL objective function

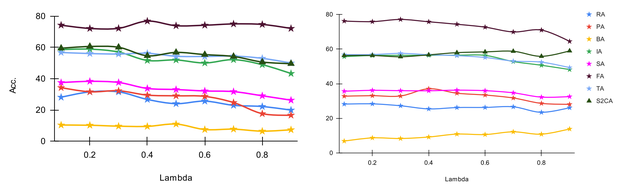
$$\mathcal{L}(\theta_e, \theta_f, \theta_d, \theta_s) = (1-\lambda) * \mathcal{L}_{\text{stutter}}(\theta_e, \theta_f, \theta_d) + \lambda * \mathcal{L}_{\text{speaker}}(\theta_e, \theta_s)$$
$$\mathcal{L}_{\text{stutter}}(\theta_e, \theta_f, \theta_d) = \mathcal{L}_{\text{fluent}}(\theta_e, \theta_f)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

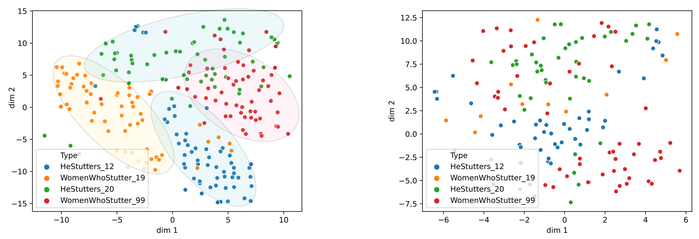
Adv Loss

$$\mathcal{L}(\theta_e, \theta_f, \theta_d, \theta_s) = \mathcal{L}_{\text{stutter}}(\theta_e, \theta_f, \theta_d) - \lambda * \mathcal{L}_{\text{speaker}}(\theta_e, \theta_s)$$
$$\mathcal{L}_{\text{speaker}} = -\sum_{(X_i, f_i, d_i, s_i)\in\mathcal{D}} \log(p(s_i/Z_i)) \quad \mathcal{L}_{\text{stutter}} = -\sum_{(X_i, f_i, d_i, s_i)\in\mathcal{D}} \log(p(d_i/Z_i)) + \log(p(f_i/Z_i)) \quad (2)$$

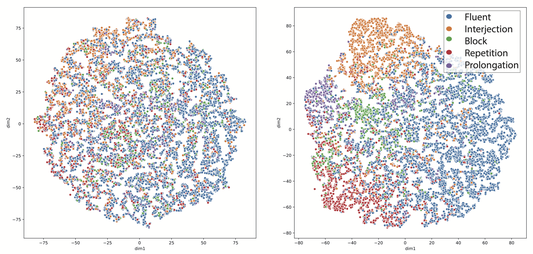


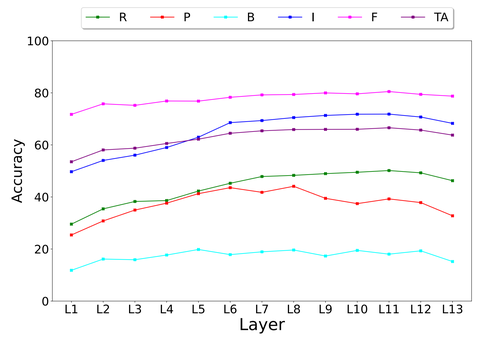






SI using Speech Embeddings





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). The results reported for Wav2Vec2.0 are from Layer 11 without fine tuning.

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 + ECAPA-TDNN + LDA	24.51	10.33	5.03	44.49	68.73	48.81
NN + Wav2Vec2.0 + LDA	50.15	39.25	18.03	71.83	80.50	66.59
NN + Fusion	44.13	38.02	18.39	66.44	84.53	67.00
NN + Layer Fusion	46.79	40.79	**23.86**	69.54	**84.32**	**68.35**

- Multi-branch (BL) scheme outperforms single branch *StutterNet* by a relative improvement of 26%.
- λ acts a control parameter for the podcast information to flow through the network.
- The well-formed podcast clusters in the MTL scheme 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 properties.
- Confusion is more among repetitions, blocks and fluent classes.
- Wav2Vec2.0 captures rich stutter discriminative features over ECAPA-TDNN.
- An overall improvement of 17.46% and 19.83% over BL using ECAPA-TDNN and Wav2Vec2.0 embedding fusion and layer fusion (L1 + L7 + L11) of Wav2Vec2.0 respectively.
- Currently working on how fine tuning of Wav2Vec2.0 can impact the SI performance.

References

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

[2] S. A. Sheikh et al. Introducing ecapa-tdnn and wav2vec2. 0 embeddings to stuttering detection. In *Proc. Interspeech 2022 (Under Review)*.

[3] S A. Sheikh et al. Stutternet: Stuttering detection using tdnn. In *Proc. 29th EUSIPCO*, 2021.

[4] S. A. Sheikh et al. Robust stuttering detection via multi-task and adversarial learning. In *Proc. 30th EUSIPCO (Under Review)*, 2022.

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