INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

A Stutter Seldom Comes Alone – Cross-Corpus Stuttering Detection as a Multi-label Problem

Sebastian P. Bayerl[1], *Dominik Wagner*[1], *Ilja Baumann*[1], *Florian Hönig*[2], *Tobias Bocklet*[1,3], *Elmar Nöth*[4], *Korbinian Riedhammer*[1]

[1] Technische Hochschule Nürnberg Georg Simon Ohm, Germany
[2]KST Institut GmbH, [3]Intel Labs
[4] Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany

sebastian.bayerl@ieee.org

Abstract

Most stuttering detection and classification research has viewed stuttering as a multi-class classification problem or a binary detection task for each dysfluency type; however, this does not match the nature of stuttering, in which one dysfluency seldom comes alone but rather co-occurs with others. This paper explores multi-language and cross-corpus end-to-end stuttering detection as a multi-label problem using a modified wav2vec 2.0 system with an attention-based classification head and multi-task learning. We evaluate the method using combinations of three datasets containing English and German stuttered speech, one containing speech modified by fluency shaping. The experimental results and an error analysis show that multi-label stuttering detection systems trained on cross-corpus and multi-language data achieve competitive results but performance on samples with multiple labels stays below overall detection results.

Index Terms: stuttering, dysfluency detection, dysfluency, cross-dataset, pathological speech

1. Introduction

Dysfluency refers to abnormalities of fluency, which includes, but is not limited to, stuttering [1]. Stuttering is a complex fluency disorder distinguished by its core symptoms, which include repetitions of words, syllables, and sounds, as well as prolongations and blocks while speaking [2]. Therapeutic options such as speech modification have been proposed to alleviate symptoms and improve fluency. Stuttering can impair a person's ability to communicate, which extends to voice technology, necessitating the detection of atypical speech and the application of custom models.

Recent work on machine learning for stuttering focused on stuttering detection [3, 4, 5] and stuttering classification [6, 7, 8, 9]. Grosz et al. used pre-trained German wav2vec 2.0 (W2V2) models and combined them with other classifiers in an ensemble approach [6]. Montacie et al. also used W2V2 features, treating them like low-level descriptors, and computed several functionals on the features, similar to the openSMILE approach [8, 10].

In [5], the authors describe an approach employing long short-term memory (LSTM) networks with a residual neural network (ResNet) backend to detect stuttering in the UCLASS corpus [11, 5]. Lea et al. applied multi-task learning with LSTMs in their dysfluency detection approach and treated stuttering detection as a multi-label problem but did not evaluate the multi-label performance [12]. They could show that dysfluency detection systems trained on one dataset generalized to another, given enough training data [12]. Recent work by Bayerl et al. showed that W2V2 feature extractors could be fine-tuned for stuttering detection using English stuttering data [3]. The respective features were transferable from English to German for all dysfluency types except word repetitions. Both contributions did not explore the effect of multi-language and cross-dataset training or evaluation on their detection systems.

The SEP-28k dataset, and the relabeled FluencyBank dataset, are fairly new resources. Unfortunately, the authors did not publish the data partitioning, making it hard to reproduce and compare results to their baseline systems [12]. Several researchers evaluated their methods but did not disclose their exact splits either, or filtered out examples that substantially impact evaluation results. Sheikh et al. removed a class by combining word- and sound repetitions and filtered out clips with an additional non-stuttering label, e.g., natural pause or background music [12], i.e., removing difficult examples. They randomly split the remaining data into a train, development, and test set (80/10/10%), irrespective of the speakers [13]. Other work filtered out all non-unanimously labeled clips from the training and test data, aiming for easy samples [14]. The work by [15] left out the block class and applied random splitting irrespective of the speaker, which can lead to overly optimistic results, as a study on the influence of dataset partitioning could show [16]. As a possible solution, the authors extended the dataset by adding semi-auto-generated speaker labels and non-speaker overlapping splits of SEP-28k, including results for baseline experiments [16].

Apart from problems with reproducibility, most works ignore that one stutter seldom comes alone, *i.e.*, multiple types of dysfluencies co-occur. In SEP-28k-Extended (SEP-28k-E), FluencyBank, and KSoF, about 23%, 29%, and 17% of clips were labeled with more than one dysfluency relevant label type – strong evidence that stuttering detection and classification are in fact multi-label problems.

Our contributions are 1) reproducible multi-label stuttering detection and classification using an end-to-end (E2E) system based on W2V2 and evaluated on three datasets, one containing modified speech (fluency shaping); 2) experimental evidence for the generalizability and feasibility of multi-language and cross-dataset training of multi-label dysfluency detection systems; 3) a detailed analysis regarding specialized binary vs. multi-label dysfluency classification that shows that classification errors happen more often on samples with multiple dysfluency labels.

2. Data

In our experiments, we use three corpora containing 3-second long clips with stuttered speech; SEP-28k-Extended, FluencyBank, and the Kassel State of Fluency (KSoF) dataset [16, 12, 17]. The SEP-28k-E contains about 28.000 English stuttered speech extracted from podcasts and is based on the SEP-28k

10.21437/Interspeech.2023-2026

corpus, extending it with speaker labels and a speaker-exclusive Train-Dev-Test split.[1]

The SEP-28k corpus contains an additional 4144 English clips extracted from the interview part of the adults who stutter dataset of the FluencyBank corpus [18] that were labeled using the same protocol. For evaluation purposes, we use the split defined by [3].[2] All three datasets were labeled similarly, containing no dysfluencies or one or more types of dysfluencies; blocks, prolongations, sound repetitions, word repetitions, and interjections. KSoF is a German dataset containing 5597 segments extracted segments from stuttering therapy recordings. In addition to the dysfluency labels in SEP-28k-E and FluencyBank, the clips were also labeled with speech modifications, marking a clip as containing a person using fluency shaping. Fluency shaping is a technique persons who stutter (PWS) learn in stuttering therapy to help them overcome their stuttering [17]. The exact label distribution of the datasets can be found in the literature [12, 17, 16]. All datasets have ambiguously labeled segments, unlike SEP-28k-E and FluencyBank; KSoF does not have samples labeled as belonging to the no dysfluencies class while at the same time being marked as containing any of the five-dysfluency labels.

To evaluate the effect of multi-lingual and multi-dataset training, we define three combinations of datasets. ALL-EN combines FluencyBank and SEP-28k-E, Multi-Lingual-Small (M-Ling-S) combines KSoF and FluencyBank, and Multi-Lingual (M-Ling) consists of KSoF, FluencyBank, and SEP-28k-E. Systems training with X-Ling-S, M-Ling, and KSoF are trained to solve a seven-class multi-label classification problem (Modified (Mod), Blocks (Bl), Interjections (Int), Prolongation (Pro), Sound Repetitions (Snd), Word Repetitions (Wd), No Dysfluencies (No-Df)). While the FluencyBank, SEP-28k-E, and ALL-EN splits are trained to detect six classes (Bl, Int, Pro, Snd, Wd, No-Df).

3. Method

3.1. wav2vec 2.0

In our experiments, we use large W2V2 models, as [6] have shown that large W2V2 models with more parameters outperform the base models with fewer parameters in the related multi-class stuttering classification task [6]. The weights for the models used in the experiments were pre-trained for English or German ASR, respectively, and are available in the referenced source code. The large W2V2 model consists of a convolutional feature extractor at the beginning of the model that takes in the waveform, followed by 24 transformer encoder blocks. The model encodes 20ms of audio into 1024-dimensional feature vectors after each transformer block, yielding 24 x t x 1024-dimensional hidden representations, with $t = \frac{time\ s}{20ms}$ [19]. W2V2 features perform well in dysfluency detection [3] and other speech tasks, such as automatic speech recognition (ASR), and mispronunciation detection [19, 20].

The model was adapted from the standard W2V2 sequence classification implementation [21] and differs mainly in three aspects: Instead of using only a single output branch, we use an alternative output branch for multi-task learning using a second output branch. Instead of using mean pooling along the time dimension, a global scaled dot-product attention mechanism takes the mean of all feature vectors along the time dimension as the query input Q, with keys K and values V being the mean over time for each dimension of the hidden states from the last hidden state [22]. This is done to better account for the different temporal contexts needed to detect each of the dysfluency classes than just using mean-pooling over time can do. The third change is a second output branch used for multi-task learning, analogous to the one described in [3]. A sigmoid function (σ) is applied to the main task outputs predicting the probability of each class for an audio clip. All models reported have an auxiliary branch with two outputs and a softmax activation function. A schematic representation of the components and changes to the default classification head is depicted in Figure 1.

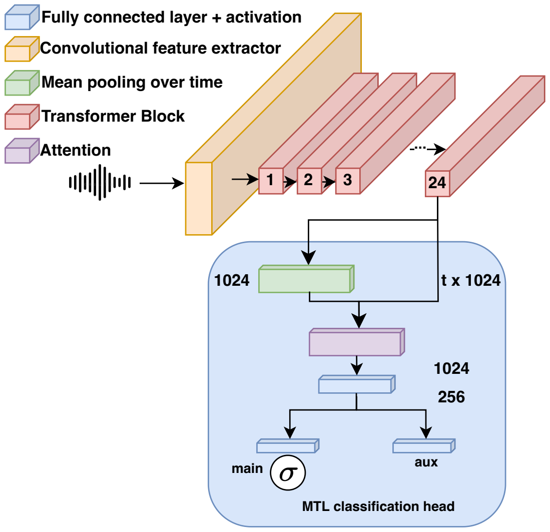

Figure 1: *Schematic overview of the wav2vec 2.0 model with modified attention-based multi-task (MTL) classification head.*

3.2. Loss

Previous studies have shown the usefulness of multi-task learning (MTL) when training dysfluency detection systems, using either an artificial ‘any’ label, indicating the presence of any dysfluency, or gender classification. [12, 3, 13].

In our experiments, we use a combination of weighted Binary Cross Entropy (BCE) and Focal Loss (FL) [23]. FL is an extension of the BCE loss using the α and γ parameters to put special emphasis on minority classes to handle class imbalance.

$$\mathbf{FL}(p_t) = -\alpha(1-p_t)^{\gamma}\log{(p_t)}. \qquad (1)$$

FL can be used equivalently to BCE loss for multi-label problems by calculating the loss for each class using the output of a given output neuron. In our experiments, the overall FL is calculated by computing the mean loss value for each class as shown in eq. (2).

$$\mathbf{FL_{multi}} = \frac{1}{n}\sum_{1}^{n}\mathbf{FL_n}(p_t) \qquad (2)$$

The final multi-task loss is computed by combining the BCE loss for the auxiliary task (L_{aux}) and FL for the main task (L_{main}) as a sum weighted by w_{main}, as shown in eq. (3).

$$\mathbf{L_{MTL}} = w_{\mathrm{main}}L_{\mathrm{main}} + (1 - w_{\mathrm{main}})L_{\mathrm{aux}} \qquad (3)$$

4. Experiments

Our experiments provide insights into dysfluency detection as a multi-label problem and the influence of training data quantity and composition across datasets and languages. Preliminary experiments included modifying the classification head of

[1] Online: https://tinyurl.com/yck9fmfv

[2] Online: https://tinyurl.com/24vm6dec

the W2V2 model by adding an attention mechanism for pooling with a trainable token parameter, mean and statistical pooling as implemented in [21], or a classification token-based mechanism as used by BERT [24], which led to slightly worse results than the employed attention-based classification head.

All experiments were performed using publicly available models of large pre-trained W2V2 feature extractors. The experiments using M-Ling, M-Ling-S, FluencyBank, SEP-28k-E, and ALL-EN training partitions were based on a model that was fine-tuned for ASR on the LibriSpeech corpus (ASR LARGE (EN)) [19, 25]. The experiments using only KSoF data for training were based on a fine-tuned model for German ASR using the Common Voice dataset (ASR LARGE (DE)).[3] To assess the effects of the multi-language and cross-corpus training, we also performed fine-tuning experiments (exp. 10 – 15), utilizing the weights obtained by training experiment 1 (cf. Table 1).

All systems were trained for up to 20 epochs using the adamW optimizer, a warm-up phase of 10% of total training steps and a patience of five epochs, choosing the best model based on the development loss [26]. During training, the convolutional feature extractor at the input of the model was frozen. The best hyper-parameters for training were determined using Optuna [27] with the tree-structured Parzen estimator (TPE) sampling strategy. The optimization algorithm could choose from an initial learning rate, $lr \in \{3 \times 10^{-5}, 3 \times 10^{-5}\}$, and batch-sizes $bs \in \{8, 32, 64, 128, 256\}$. The main loss weight w_{main} and the FL parameters α and γ were determined from $w_{main} \in \{0.85, ...0.95\}$, $\gamma \in \{1, 2, 3\}$, and $\alpha \in \{0.1, 0.2, \ldots, 0.9\}$. The auxiliary task was chosen from either gender classification or an artificial “any” label, indicating the presence of any dysfluency in a segment, for experiments using only a single dataset and the ALL-EN combination. Experiments using M-Ling-S and M-Ling data also used language-id as their potential auxiliary task. The best-performing parameters across all experiments were $lr = 3 \times 10^{-5}$, $bs = 8$, $\alpha = 0.7$, and $\gamma = 3$. The best performing auxiliary task for experiments using only the SEP-28k-E and ALL-EN data was detecting any dysfluency with $w_{\text{main}} = 0.9$. For KSoF and FluencyBank, gender classification led to the best overall results with $w_{\text{main}} = 0.92$. Training using the M-Ling and M-Ling-S data profited most from the language-id task with $w_{\text{main}} = 0.92$.

5. Results and Discussion

This section describes the overall and specific results from Table 1 and Table 2. All systems are evaluated on their training data. Their cross-dataset and cross-language performance is tested by evaluating the models on the corpora that were not used for training. We report F1 scores for all dysfluency types and modifications. The result N/A indicates cases where precision and recall are zero per definition, as there are no labeled clips in the respective partition, i.e., F1 is undefined. The results were balanced between *precision* and *recall*. Table 3 considers multi-label classification metrics for each dataset.

Table 1 shows the effect of training data quantity.The system trained on the larger SEP-28k-E dataset generalizes well to FluencyBank, improving or matching results reported in the literature with expert binary dysfluency detection systems (Baseline systems), unlike the system trained using only FluencyBank which generalizes poorly to SEP-28k-E and KSoF. The multi-label systems trained on SEP-28k-E or KSoF, improve or match the baseline results for all dysfluency classes except **Wd** for KSoF and **Bl** for SEP-28k-E.

Table 1: *Multi-label dysfluency detection results (F1-score) using multi-label E2E systems. (**Mod** = Modified Speech, **Bl** = Block, **Int** = Interjection, **Pro** = Prolongation, **Snd** = Sound repetition, **Wd** = Word repetition). Section headers indicate the training data, followed by the W2V2 weights used.*

#	Test	Mod	Bl	Int	Pro	Snd	Wd
Baseline Systems							
-	SEP-28k-E [16]	-	0.33	0.68	0.46	0.39	0.51
-	FBANK [3]	-	0.33	0.84	0.60	0.60	0.43
-	KSoF [3]	0.76	0.54	0.74	0.53	0.47	0.19
SEP-28k-E (ASR LARGE EN)							
1	**SEP-28k-E**	-	0.16	**0.77**	**0.51**	**0.50**	**0.62**
2	FBANK	-	0.17	**0.82**	0.60	**0.63**	**0.47**
3	KSoF	-	0.10	0.55	0.44	0.35	**0.23**
FluencyBank (ASR LARGE EN)							
4	SEP-28k-E	-	0.00	0.65	0.41	0.42	0.16
5	**FBANK**	-	0.00	0.73	0.44	0.55	0.25
6	KSoF	-	0.00	0.46	0.35	0.38	0.06
KSoF (ASR LARGE DE)							
7	SEP-28k-E	N/A	0.27	0.55	0.40	0.42	0.11
8	FBANK	N/A	0.27	0.60	0.56	0.58	0.12
9	**KSoF**	**0.76**	**0.60**	**0.88**	**0.57**	0.48	0.18
FluencyBank (Stutter LARGE EN)							
10	*SEP-28k-E*	NaN	0.22	0.72	0.49	0.46	0.64
11	**FBANK**	NaN	0.24	0.79	**0.61**	0.60	0.51
12	KSoF	NaN	0.30	0.59	0.42	0.49	0.19
KSoF (Stutter LARGE EN)							
13	*SEP-28k-E*	NaN	**0.30**	0.67	0.39	0.42	0.25
14	FBANK	NaN	**0.29**	0.66	0.51	0.57	0.11
15	**KSoF**	0.65	0.49	0.79	0.53	**0.51**	0.11

The fine-tuning experiments using KSoF data mostly help to improve the results in the cross-dataset setting substantially but otherwise fail to improve results over the systems based on the ASR LARGE (DE) model by a large margin. Results for FluencyBank improve substantially over the system trained using FluencyBank and the ASR LARGE (EN) model.

A comparison of Table 2 and Table 1 reveals that the models trained on the multi-language and multi-dataset data (M-Ling-S, M-Ling) achieve the overall best results, except for **Int** on FluencyBank and **Wd** on KSoF (exp. 19–24). The models achieve the best per-dysfluency performance in most cases, even though the models trained on the multi-language data have to account for the additional **Mod** class, showing their potential to generalize across languages.

The model trained using KSoF data achieved the highest F1 for **Mod** (exp. 9), which is matched by the model trained using the M-Ling-S data. Performance decreased slightly in the model trained with M-Ling data (exp. 21, 24). Evaluating the recognition of **Mod** with the multi-lingual models shows no confusion for **Mod** to any English segments and vice-versa. This might be due to the language or the distinctiveness of the pattern, but also, different recording conditions may cause these results.

To evaluate the multi-label performance, we use the exact match ratio (EMR), the partial match ratio (PMR), recall, and the Hamming loss (HL). The EMR is a pessimistic evaluation metric, extending accuracy to the multi-label case [28], ignoring the notion of partially correct samples. The PMR is the ratio

[3] Code available: `https://tinyurl.com/mr4h3ybk`

Table 2: *Cross-Dataset and multi-Lingual dysfluency detection results (F1-score) using E2E multi-label systems. (**Mod** = Modified Speech, **Bl** = Block, **Int** = Interjection, **Pro** = Prolongation, **Snd** = Sound repetition, **Wd** = Word repetition). Headers indicate the training data, followed by the W2V2 weights used.*

#	**Test**	**Mod**	**Bl**	**Int**	**Pro**	**Snd**	**Wd**
ALL-EN (ASR LARGE EN)							
16	SEP-28k-E	-	0.22	0.74	0.51	0.49	0.63
17	FBANK	-	0.23	**0.80**	0.59	0.63	0.51
18	KSoF	-	0.35	0.65	0.37	0.47	**0.18**
Multi-Lingual-Small (ASR LARGE EN)							
19	SEP-28k-E	N/A	0.25	0.70	0.50	0.45	0.26
20	FBANK	N/A	0.09	0.75	0.56	0.61	0.40
21	KSoF	**0.76**	0.56	**0.90**	0.58	**0.54**	0.11
Multi-Lingual (ASR LARGE EN)							
22	SEP-28k-E	N/A	**0.32**	**0.77**	**0.53**	**0.53**	**0.64**
23	FBANK	N/A	**0.36**	0.79	**0.62**	**0.64**	**0.52**
24	KSoF	0.75	**0.64**	0.85	**0.60**	0.48	0.14

of samples for which at least one label was correctly recalled to the total number of samples. The HL specifies the fraction of incorrectly classified labels; it considers both the prediction error and the missing error normalized by the total number of classes and examples [28].

The results in Table 3 and the following analysis consider the overall best-performing model trained on the M-Ling data. The EMR results for all three datasets are around 0.5. Only looking at the segments that have more than one label drops these results to around 0.2, indicating that many classification errors occur in segments that have multiple labels. The increased HL indicates that the model makes about twice as many mistakes on segments with multiple labels. In our error analysis of the multi-label results, we, therefore, looked at the most common combinations of segments labeled with two types of dysfluencies for each dataset and picked three label combinations for discussion. To be included in the analysis, there had to be at least 50 segments for each combination in the dataset. Results are contained in Table 4

Int and **Wd** are the most frequent combination of two dysfluencies in SEP-28k-E and FluencyBank. The results show a higher EMR in the multi-label case for this label combination than for the overall dataset, with a very high partial match rate. **Ints** are generally recognized well across all systems, and the model tends to favor its prediction. In the partial matches, the prediction of **Int** dominates in cases where the model only predicts one of the labels correctly. The predictions for the combination of **Int** and **Snd** are similar, with higher EMR values for the label combination than the overall dataset and a high PMR value. Still, **Ints** are recalled more often than **Snd**. Again, the model favors predicting only **Ints**, which seems to hinder the prediction of **Snd**. These co-occurrences and the favored prediction of **Ints** have implications for the evaluation of binary detection systems, where it is possible that the model rather learns to recognize the **Ints** in the segments rather than the target dysfluency or is distracted by the most dominant pattern. This can influence results and has to be accounted for during training time if the pattern co-occurs often enough.

The combination of **Pro** and **Bl** has a zero EMR for SEP-28k-E. The model fails to predict the combination completely and also has a very low EMR for KSoF. The models have difficulties detecting **Bl** on their own (see Bl results in Table 2) and even more so in the multi-label case. Assessing **Bl** using only acoustics is a hard task, even for clinicians that usually rely on signs of physical tension and grasping for air when assessing blocks. This is reflected by low inter-rater reliability (IRR)(Fleiss κ, 0.25 SEP-28k, 0.37 KSoF) for **Bl** reported for the datasets [12, 17]. Only the results for KSoF manage to improve over previous systems (Baseline) and manage to recall 42% of **Bl** in the multi-label case (see Table 4). We hypothesize that this is due to KSoF consisting of therapy recordings that include PWS who, on average, have more pronounced symptoms, which is supported by the higher κ.

The analysis of multi-label errors shows the necessity for measures to deal with multi-label dysfluency data during training time. Possible measures might include modified loss functions and the inclusion of prior knowledge about common co-occurring dysfluencies.

Table 3: *Hamming Loss (HL) and Equal Match Ratio (EMR) for the test set of each dataset (Test) and segments from the test set that were labeled with more than one dysfluency (Test-Multi)*

	Test		**Test Multi**	
Dataset	**HL**	**EMR**	**HL**	**EMR**
SEP-28k-E	0.13	0.54	0.23	0.2
FBANK	0.11	0.53	0.21	0.23
KSoF	0.11	0.49	0.18	0.21

Table 4: *Equal match ratio, partial match ratio (PMR), and recall (Re) for the first (L1) and second label (L2).*

L1 & L2	Dataset	EMR	PMR	Re L1	Re L2
Int & **Wd**	SEP-28k-E	0.44	0.95	0.79	0.60
Int & **Wd**	FluencyBank	0.39	0.97	0.88	0.48
Int & **Snd**	SEP-28k-E	0.36	0.90	0.74	0.51
Int & **Snd**	FluencyBank	0.33	0.87	0.80	0.40
Int & **Snd**	KSoF	0.29	0.93	0.79	0.43
Pro & **Bl**	SEP-28k-E	0.00	0.76	0.69	0.07
Pro & **Bl**	KSoF	0.16	0.84	0.58	0.42

6. Conclusion

This paper introduced a multi-label W2V2-based end-2-end stuttering detection and classification system. Leveraging multi-language training data from multiple datasets, previous results achieved by binary per-dysfluency classification systems could, for the most part, be improved on each of the FluencyBank, SEP-28k-E, and KSoF datasets, even though multi-label detection is a more complex task.

Furthermore, results emphasize the difficulty of co-occurring dysfluencies when training binary-detection systems. For multi-label systems, these patterns should be accounted for during training time to improve detection results and prevent errors on test samples with multiple labels. The experimental results show that the quantity and diversity of training data are still the most important factors for creating more robust dysfluency detection systems, even if using large pre-trained models such as W2V2. This is even more true in the case of multi-label stuttering detection, where there are only a few samples for each dysfluency combination.

In future work, we will therefore explore prior-knowledge-infused handling of segments with co-occurring dysfluencies as well as data augmentation techniques to increase or improve the respective classification performance.

7. References

[1] M. E. Wingate, “Fluency, disfluency, dysfluency, and stuttering,” *Journal of Fluency Disorders*, vol. 9, no. 2, pp. 163–168, May 1984.

[2] R. Lickley, “Disfluency in typical and stuttered speech,” *Fattori sociali e biologici nella variazione fonetica*, no. 3, p. 373, 2017.

[3] S. P. Bayerl, D. Wagner, E. Noeth, and K. Riedhammer, “Detecting Dysfluencies in Stuttering Therapy Using wav2vec 2.0,” in *Proc. Interspeech 2022*. ISCA, Sep. 2022, pp. 2868–2872.

[4] J. Harvill, M. Hasegawa-Johnson, and C. D. Yoo, “Frame-level stutter detection,” in *Proc. Interspeech 2022*, 2022, pp. 2843–2847.

[5] T. Kourkounakis, A. Hajavi, and A. Etemad, “Detecting Multiple Speech Disfluencies using a Deep Residual Network with Bidirectional Long Short-Term Memory,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6089–6093.

[6] T. Grósz, D. Porjazovski, Y. Getman, S. Kadiri, and M. Kurimo, “Wav2vec2-based paralinguistic systems to recognise vocalised emotions and stuttering,” in *Proceedings of the 30th ACM International Conference on Multimedia*, ser. MM ’22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 7026–7029.

[7] S. A. Sheikh, M. Sahidullah, S. Ouni, and F. Hirsch, “End-to-end and self-supervised learning for ComParE 2022 stuttering sub-challenge,” in *Proceedings of the 30th ACM International Conference on Multimedia*, ser. MM ’22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 7104–7108.

[8] C. Montacié, M.-J. Caraty, and N. Lackovic, “Audio features from the Wav2Vec 2.0 embeddings for the ACM multimedia 2022 stuttering challenge,” in *Proceedings of the 30th ACM International Conference on Multimedia*, ser. MM ’22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 7195–7199.

[9] K. You, K. Xu, B. Zhu, M. Feng, D. Feng, B. Liu, T. Gao, and B. Ding, “Masked modeling-based audio representation for ACM multimedia 2022 computational paralinguistics ChallengE,” in *Proceedings of the 30th ACM International Conference on Multimedia*, ser. MM ’22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 7060–7064.

[10] F. Eyben, M. Wöllmer, and B. Schuller, “Opensmile: The munich versatile and fast open-source audio feature extractor,” in *Proceedings of the International Conference on Multimedia - MM ’10*. Firenze, Italy: ACM Press, 2010, p. 1459.

[11] P. Howell, S. Davis, and J. Bartrip, “The University College London Archive of Stuttered Speech (UCLASS),” *Journal of Speech, Language, and Hearing Research*, vol. 52, no. 2, pp. 556–569, Apr. 2009.

[12] C. Lea, V. Mitra, A. Joshi, S. Kajarekar, and J. P. Bigham, “SEP-28k: A Dataset for Stuttering Event Detection from Podcasts with People Who Stutter,” in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Toronto, ON, Canada: IEEE, Jun. 2021, pp. 6798–6802.

[13] S. A. Sheikh, F. Hirsch, and S. Ouni, “Robust Stuttering Detection via Multi-task and Adversarial Learning,” in *2022 30th European Signal Processing Conference (EUSIPCO)*, 2022, p. 5.

[14] P. Mohapatra, A. Pandey, B. Islam, and Q. Zhu, “Speech disfluency detection with contextual representation and data distillation,” in *Proceedings of the 1st ACM International Workshop on Intelligent Acoustic Systems and Applications*, ser. IASA ’22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 19–24.

[15] M. Jouaiti and K. Dautenhahn, “Dysfluency Classification in Stuttered Speech Using Deep Learning for Real-Time Applications,” in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Singapore, Singapore: IEEE, May 2022, pp. 6482–6486.

[16] S. P. Bayerl, D. Wagner, E. Nöth, T. Bocklet, and K. Riedhammer, “The Influence of Dataset Partitioning on Dysfluency Detection Systems,” in *Text, Speech, and Dialogue*, P. Sojka, I. Kopeček, K. Pala, and A. Horák, Eds. Springer International Publishing, 2022.

[17] S. P. Bayerl, A. Wolff von Gudenberg, F. Hönig, E. Noeth, and K. Riedhammer, “KSoF: The Kassel State of Fluency Dataset – A Therapy Centered Dataset of Stuttering,” in *Proceedings of the Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, Jun. 2022, pp. 1780–1787.

[18] N. Bernstein Ratner and B. MacWhinney, “Fluency Bank: A new resource for fluency research and practice,” *Journal of Fluency Disorders*, vol. 56, pp. 69–80, Jun. 2018.

[19] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations,” in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 12 449–12 460.

[20] X. Xu, Y. Kang, S. Cao, B. Lin, and L. Ma, “Explore wav2vec 2.0 for Mispronunciation Detection,” in *Interspeech 2021*. ISCA, Aug. 2021, pp. 4428–4432.

[21] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, “Transformers: State-of-the-Art Natural Language Processing,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is All you Need,” in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[23] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, “Focal Loss for Dense Object Detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *arXiv:1810.04805 [cs]*, May 2019.

[25] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An ASR corpus based on public domain audio books,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. South Brisbane, Queensland, Australia: IEEE, Apr. 2015, pp. 5206–5210.

[26] I. Loshchilov and F. Hutter, “Decoupled Weight Decay Regularization,” *arXiv:1711.05101 [cs, math]*, Jan. 2019.

[27] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, “Optuna: A Next-generation Hyperparameter Optimization Framework,” in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Anchorage AK USA: ACM, Jul. 2019, pp. 2623–2631.

[28] M. S. Sorower, “A literature survey on algorithms for multi-label learning,” 2010.