

# KSoF: The Kassel State of Fluency Dataset – A Therapy Centered Dataset of Stuttering

Sebastian P. Bayerl<sup>1</sup>, Alexander Wolff von Gudenberg<sup>3</sup>, Florian Hönig<sup>3</sup>,  
Elmar Nöth<sup>2</sup>, and Korbinian Riedhammer<sup>1</sup>

<sup>1</sup>Technische Hochschule Nürnberg Georg Simon Ohm

<sup>2</sup>Friedrich-Alexander Universität Erlangen-Nürnberg

<sup>3</sup>Institut der Kasseler Stottertherapie

sebastian.bayerl@ieee.org

## Abstract

Stuttering is a complex speech disorder that negatively affects an individual’s ability to communicate effectively. Persons who stutter (PWS) often suffer considerably under the condition and seek help through therapy. Fluency shaping is a therapy approach where PWSs learn to modify their speech to help them to overcome their stutter. Mastering such speech techniques takes time and practice, even after therapy. Shortly after therapy, success is evaluated highly, but relapse rates are high. To be able to monitor speech behavior over a long time, the ability to detect stuttering events and modifications in speech could help PWSs and speech pathologists to track the level of fluency. Monitoring could create the ability to intervene early by detecting lapses in fluency. To the best of our knowledge, no public dataset is available that contains speech from people who underwent stuttering therapy that changed the style of speaking. This work introduces the Kassel State of Fluency (KSoF), a therapy-based dataset containing over 5500 clips of PWSs. The clips were labeled with six stuttering-related event types: blocks, prolongations, sound repetitions, word repetitions, interjections, and – specific to therapy – speech modifications. The audio was recorded during therapy sessions at the *Institut der Kasseler Stottertherapie*. The data will be made available for research purposes upon request.

**Keywords:** stuttering, pathological speech, corpus, speech resource

## 1. Introduction

Stuttering is a complex speech disorder that affects about 1 % of people (Yairi and Ambrose, 2013). It can be identified by an increased duration and occurrence of dysfluencies, such as repetitions, prolongations of sounds, syllables or words, and blocks while speaking (Lickley, 2017). These so-called core symptoms are often accompanied by various linguistic, physical, behavioral, and emotional symptoms. Stuttering symptoms and severity vary greatly between different speakers and within the same speaker. The unique appearance and severity of stuttering symptoms are influenced by the communication situation, psychological factors, the linguistic complexity of an utterance, and the typical phased progress of the speech disorder (Ellis and Ramig, 2009). The ability to communicate can be severely disturbed and thereby negatively affect the life of a person who stutters (PWS). Besides the high remission rate during adolescence, the condition is not curable but treatable.

There are several treatment options available that aim at different goals. A common goal of therapy approaches is to increase the communication ability. Some therapy approaches aim at increasing speech fluency (Ingham et al., 2015). In contrast, others try to make people change their attitude towards their stuttering and primarily target the psychological side-effects of stuttering (Mongia et al., 2019), while other approaches use a mix of the previously mentioned methods (Euler et al., 2009).

Before or during therapy, the need to assess stuttering

severity arises, therefore, a speech pathologist quantifies dysfluency events and types before recommending therapeutic measures. This is usually done during therapy sessions where PWSs perform specific speech tasks such as reading, dialogues, or scene description. The evaluation of such tasks is highly subjective and only considers one type of communication situation, namely the therapeutic situation. It can be shown that the use of popular evaluation metrics, such as percent stuttered syllables (%SS), is not reliable to detect changes in one speaker if judged by only a single rater (Karimi et al., 2014).

An overall assessment in different communication situations would be ideal but laborious, and if only performed by a single therapist, would not remove subjectivity. Realistically only the continuous automatic detection of stuttering and dysfluency symptoms inside the home of PWS or their workplace would unlock objective assessment of stuttering. Such an assessment would enable speech therapists to plan a tailored therapy that fits the PWS’s needs by using the additional information created by the assessment of the PWS speech in various communication situations. Automatic evaluation does not only allow for a better initial assessment. Since stuttering therapy has high relapse rates it would benefit from speech monitoring. It would allow the construction of an early-warning system that enables the PWS and the speech therapist to act on the collected data and decide on further therapeutic measures. As many PWS have learned some speech techniques to overcome their stuttering, the usage of such would also have to be detected

reliably.

Our contributions are:

- Collection and annotation of a new therapy centered dataset containing German stuttered speech with six types of typical stuttering symptoms marked, including speech modifications.
- Baseline machine learning experiments for the detection of five types of stuttering as well as modified speech.<sup>1</sup>
- Insights into improving the reliability of stuttering annotations when working with naive listeners.
- Compatible dysfluency labels to the biggest publicly available resource containing stuttered speech, enabling cross-language transfer learning.

## 2. Related Work and Data

One of the main issues with creating reliable systems that can automatically detect stuttering in speech is too little data that captures the considerable variance in stuttering. Most datasets are either small, not publicly available, or labeled differently, making it difficult to compare results or transfer knowledge.

(Noeth et al., 2000) used a non-public dataset of read speech with balanced classes consisting of 37 speakers and 52 recordings. (Świetlicka et al., 2013) used artificial neural networks (ANN) to detect three types of stuttering on a non-public dataset containing 19 speakers performing a description and a reading task. A much-cited resource for the automatic detection of stuttering from speech is the University College London Archive of Stuttered Speech (UCLASS) (Howell et al., 2009). (Kourkounakis et al., 2020) created labels for a subset of the UCLASS corpus but did not publish the annotations. (Riad et al., 2020) used the adults who stutter (AWS) subset of the Fluency Bank corpus and created annotations for two speech and language tracks (primary and collateral), not considering blocks, prolongations, and syllable repetitions. The LibriStutter dataset is a synthesized dataset created based on the public LibriSpeech corpus containing labels for five types of stuttering dysfluencies (Kourkounakis et al., 2021).

A recent effort to solve the problem of data scarcity is the Stuttering Events in Podcasts (SEP-28k) dataset. It consists of speech clips extracted from podcasts from and with PWS that focus on stuttering, making it by far the largest publicly available resource on stuttered speech. In addition to labeling the podcast data, they created compatible labels for Fluency Bank to make results easier to compare (Lea et al., 2021; Bernstein Ratner and MacWhinney, 2018).

---

<sup>1</sup>The annotated data will be made available to researchers upon request; please contact [korbinian@ieee.org](mailto:korbinian@ieee.org)

None of these datasets contain speech marked as using a fluency enhancing technique, as people are taught in stuttering therapy such as fluency shaping or modified phonation intervals (MPI). (Świetlicka et al., 2013) even explicitly asked people not to use fluency enhancing techniques therapy in their recordings. Reliably detecting modified speech is important, as it enables the automatic assessment of stuttering in people who already underwent stuttering therapy or use such methods in a therapeutic context. It allows to correctly attribute a potential gain in fluency to the speech technique.

Fluency typically improves throughout therapy (Euler et al., 2009; Bayerl et al., 2020). However, it is not easy to assess the level of fluency and the adoption of speech technique of a PWS after the end of therapy. Besides regular appointments, people are on their own, and no objective measure of therapy success is available. People under supervision, i.e., in a therapy environment, act or are perceived differently; therefore, the appointments can only give a snapshot of the actual performance (Porta et al., 2014, p. 127, 205). Monitoring everyday adoption of speech techniques and speech fluency can provide important insight to therapists to make informed, data-driven decisions when it comes to exercises. It can also be used to give feedback and encourage the PWS.

## 3. Kassel State of Fluency

Speech therapy has the goal to improve speech skills of people with speech and language disabilities. In the case of stuttering therapy, fluency is not necessarily the primary goal, but getting back a sense of control over one's speech. One of the assumptions behind learning a speaking technique is that it is better to talk "funny" instead of not talking at all or saying something unintended. It takes time and effort to learn a new way of speaking that goes beyond the duration of therapy. At the Kasseler Stottertherapie (KST), therapy is split in three stages. The first stage is an initial assessment and a discussion of goals with the client. The second stage is a two-week full-time (on-premises) intensive course in which participants learn a new speech technique and train in real-life situations, such as shopping in a bakery or flower shop or calling somebody unknown on the phone. The third stage follows the intensive course for one year. Clients use self-directed learning with the help of an online tool and occasional therapy sessions. All recordings in this dataset were created during these three stages of therapy.

The recordings in this dataset contain three types of tasks, which are spontaneous speech (SPO), reading (REA), and telephone conversations (PHO). SPO can be any open communication situation, like ordering at the bakery or speaking about therapy success with the therapist. REA is a relatively controlled task where clients were asked to read a paragraph from a given text. PHO is a planned but spontaneous speech task

Stuttering Labels	KSoF	SEP-28k	Description
Block	20.74 %	12.0 %	Gasps for air or stuttered pauses
Prolongation	12.02 %	10.0 %	Elongated syllable or Sound “[IIII]”, otherwi[ssss]se
Sound Repetition	14.76 %	8.3 %	Repeated syllables “[nat-nat-nat-naturally]” or sounds “I [t-t-t]talked to dad.
Word / Phrase Repetition	3.88 %	9.8 %	“I have [I have] done no such thing”
No dysfluencies	24.75 %	56.9 %	There are no audible dysfluencies
Modified/ Speech technique	24.44 %	- %	Soft voice onset, at the start of syllables, voluntary prolongation with continuous phonation e.g., rrReading, prrooolongation
Interjection	12.97 %	21.2 %	Filler words e.g., “ähm”, “äh”, “naja”, eng: “uhm”, “uh”
Non Stuttering Labels			
Natural pause	1.97 %	8.5 %	A non-stuttered, significant pause in speech
Unintelligible	2.00 %	3.7 %	The speech is difficult to understand
Unsure	0.30 %	0.1 %	An annotator was unsure of their response
No Speech	0.39 %	1.1 %	The clip contains no speech or is silent
Poor Audio Quality	0.98 %	2.1 %	There are microphone or other quality issues
Music (Background Noise)	0.13 %	1.1 %	Audible noise or music playing in the background

Table 1: Distribution of annotations of 3 second segments in the Kassel State of Fluency (KSoF) dataset where at least two annotators applied a given label. SEP-28k label distribution for reference (Lea et al., 2021).

involving cold-calling unacquainted people for inquiry purposes over the phone. For example, clients were asked to retrieve hotel booking information or opening hours and prices from a swimming pool. Clients can prepare in advance but have to spontaneously deal with the reactions of their dialogue partner, who is unaware that a PWS is calling. These exercises help the PWS to grow accustomed to the speech technique and deal with possible adverse reactions of dialogue partners.

The KSoF dataset can be used to train systems that recognize speech techniques learned to overcome stuttering. This can help improve therapy by enabling data-driven therapeutic decisions-making that includes speech techniques. Interventions can be made, and exercises recommended when needed and not only when scheduled.

### 3.1. Recordings

This dataset consists of clips extracted from 214 recordings by 37 speakers, of which 28 were male and 9 female, containing stuttered speech. The language spoken throughout all clips is – more or less regionally accented – German. The gender distribution matches the general ratio of males to females in stutterers, where about four times as many males stutter. The recordings were created using either a voice recorder with a close-talking microphone or the audio was extracted from video recordings created at the initial therapy sessions. The audio data was downsampled to 16 kHz and converted to one channel. The data was anonymized by removing mentions of individuals’ names. These sections in the audio signal were set to zero to ensure the participants’ privacy.

### 3.2. Annotation

The recordings described in section 3.1 had originally been annotated with an event-based approach, while also marking the exact time spans from beginning to the end of dysfluency events (Valente et al., 2015; Bayerl et al., 2020). The annotation was performed by speech therapists that had experience with stuttering therapy and the speech technique taught at KST. Unfortunately, such accurate event-based annotation proved impractical to label large amounts of data with multiple annotators since it is time consuming and requires experienced speech therapists to perform the annotation. The huge amount of English labeled stuttering data available through the work done by (Lea et al., 2021) motivated us to employ a similar time-interval based annotation approach (Valente et al., 2015). A segment length of 3 seconds seems a reasonable compromise between a satisfactory level of agreement and the structure of dialogue and spontaneous speech (Cordes and Ingham, 1994).

Our approach primarily differs regarding to annotator training and the annotation tool used. The changes implemented were supposed to lead to a better agreement among annotators and thus to a better quality of the resulting labels while keeping compatibility to the large corpus to enable easy transfer learning.

Before starting the annotation process described in this paper, the recordings were manually segmented to utterances while excluding back-channels or answers of dialogue partners. These manually generated segments were then automatically split into 3-second long segments with 1.5-second overlap, resulting in 5597 clips. All clips were annotated by three annotators. The annotators were graduate and undergraduate students from

non-speech and non-health-related studies. Prior to this task, non of the annotators had previous experience with labeling data for machine learning and had no previous prolonged exposure to PWS. They can therefore be assumed to be naive listeners.

The annotators were given a short 30-minute introduction to stuttering, stuttering therapy, and the labeling tool. In addition to the introduction, written labeling guidelines that included listening examples were provided to them. Together with the listening examples, the guidelines were accessible to them during the annotation process. Annotators were asked to mark all stuttering- and non-stuttering-labels they could identify in the recording. All labels were designed as binary choices. Label types, the resulting label distribution, and a short description of the label can be found in Table 1.

Figure 1 shows a screenshot of the browser-based online annotation tool. The interface featured a large audio player at the top that displays the waveform and can be used for navigation in the recording, and always displays the current location of the audio being played. Buttons for starting, pausing, and stopping the recording are located below the waveform plot. Annotators could listen to the sample as many times as needed. A tooltip was displayed when hovering over one of the label options to make the annotation task easier. After submitting the annotations for a clip, a new clip was randomly sampled from the remaining unlabeled clips.

The annotation process took place in two stages: the test and the main stages. The test stage started with a short labeling task to check the agreement between the annotators and see if there was a common understanding or misconceptions about the annotation task itself or the labels. Annotators had to annotate 123 clips from four subjects one female and three male. After the initial annotation task, the test stage concluded with a one-hour-long meeting with all annotators. During the meeting, samples with a low agreement and prominent features for specific stuttering symptoms were discussed. In the main stage of the annotation process, annotators had to assign labels to all clips in the dataset. They were on their own, did not communicate with each other, and only had access to the written annotation guidelines and the listening examples.

Table 2 contains Fleiss’ kappa agreement metrics for all stuttering-related label types. The initial agreement for prolongations was very poor and improved to a moderate agreement in the main task (Landis and Koch, 1977). To our surprise, interjections had the second-lowest agreement in the test stage and improved to a substantial agreement. The agreement for no dysfluency, word repetition, and sound repetition improved slightly, whereas the value for modified stayed the same.

Overall agreement is higher than expected which can be seen by comparing values from agreement values

Stuttering Labels	KSoF	Test	Sep28k
Block	0.37	0.60	0.25
Prolongation	0.42	0.06	0.11
Sound Repetition	0.54	0.52	0.40
Word / Phrase Repetition	0.59	0.57	0.62
No dysfluencies	0.59	0.40	0.39
Interjection	0.78	0.23	0.57
Modified/ Speech technique	0.55	0.55	-

Table 2: Fleiss’ kappa agreement statistics for each type of stuttering. Table contains values for the test labeling task (Test), the overall task (KSoF); SEP-28k agreement for comparison (Lea et al., 2021).

achieved in a similar annotation task which can be seen in Table 2.

It can be hypothesized that the resulting higher-than-expected agreement was caused by the spoken instructions and introduction to the topic instead of just written guidelines and also the building of a shared understanding during the agreement meeting at the end of the test stage. The general agreement in the test stage and the only fair agreement for blocks in the main stage underline the difficulty of the annotation task.

### 3.3. Metadata

Metadata complement the annotations. They can help with error analyses and the creation of new experiments and views of the data. For each clip, we provide information about gender, therapy status, type of microphone used, task performed by the PWS, original unique recording the clip was extracted from, and speaker.

### 3.4. Suggested Evaluation

Evaluation of health conditions in small datasets generally comes with some challenges. Specifically for stuttering, there is a considerable inter- and intra-*speaker* variance of stuttering behavior that is dependent of factors such as the communication situation, psychological factors, and the linguistic complexity of an utterance (Ellis and Ramig, 2009). A small dataset can hardly capture this variance.

Providing no fixed data partitioning can lead to cherry-picking and, hence, to overly optimistic results that are not reproducible, leading to an unrealistic view of the transferability of results. The best possible evaluation would therefore use a leave-one-speaker out approach. For KSoF, this would require training 37 models with every experiment, which seems impractical, especially when training times are long and resources are limited. For small datasets such as KSoF, we believe that a speaker disjoint k-fold cross-validation (CV) is a good compromise to ensure generalization and objective results. This evaluation strategy has its pitfalls, such as improper class- or gender distribution among folds. Even the distribution of the communication situations might influence on fold-performance.

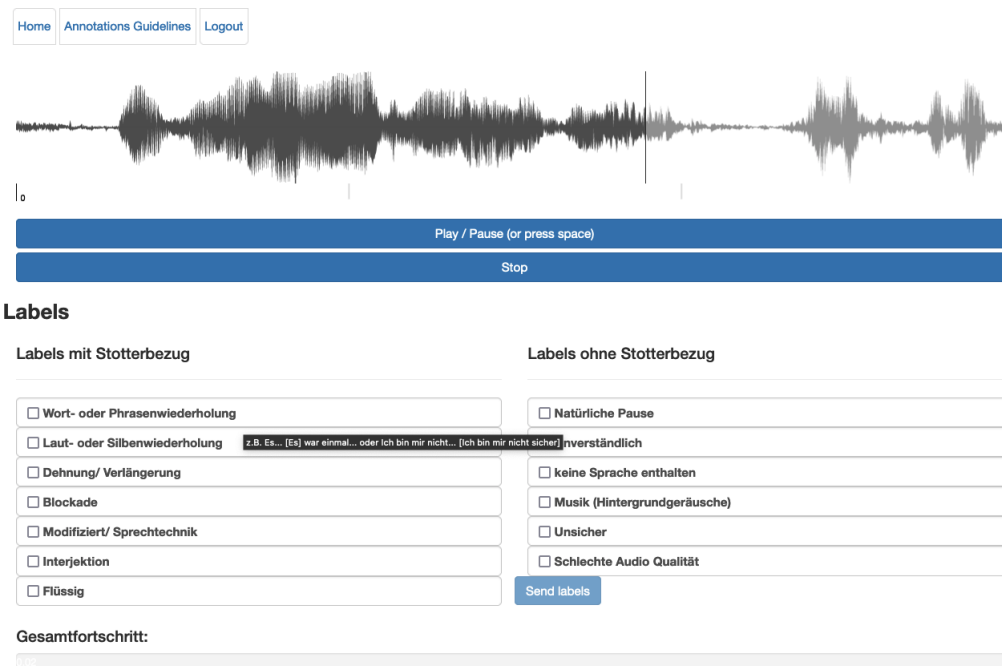


Figure 1: Screenshot of the annotation tool displaying an audio player with a plot of the audio wave-form that can be used for navigating the audio. Labels were arranged in two columns: stuttering related labels and non-stuttering related labels. A tooltip describing the label is being displayed upon hovering the mouse over the label.

At the same time, we recognize the community’s need for a simple and easily comparable data split into training, validation and test set. The dataset therefore includes a suggested split for easy and quick comparison. For KSoF, we strongly recommend using at least five-fold cross-validation when working with the data.

## 4. Methods

In the following, we briefly introduce the methods used to compute a variety of baseline classification experiments.

### 4.1. openSMILE

As our initial baseline, we chose to use the openSMILE toolkit to extract the ComParE 2016 feature set consisting of 6373 static features from the computation of various functionals over low-level descriptor (LLD) contours (Schuller et al., 2016). OpenSMILE features are widely used and have been shown to achieve proper baseline performance in numerous paralinguistic applications such as gender detection, age detection, or speech emotion recognition (Schuller et al., 2016; Schuller et al., 2021).

We trained a Support Vector Machine (SVM) with a Gaussian kernel on the openSMILE features. Before training the SVM, we performed a principal component analysis (PCA) to reduce the negative effect of highly correlated features on most classification systems. We transformed each openSMILE feature vector to a 100-dimensional vector per clip.

### 4.2. wav2vec 2.0

Neural networks benefit from large quantities of labeled training data. Suppose this labeled in-domain data is not available in sufficient quantities. In this case it is common to use models trained on large amounts of related data as feature encoders, as they have learned latent representations describing many aspects of the underlying data. The wav2vec 2.0 (W2V2) approach learns a set of speech units from large amounts of data. W2V2 mainly consists of a convolutional neural network (CNN) encoder, a contextualized transformer network, and a quantization module. It takes raw wave-files as inputs, and the CNN produces latent representations that the quantization module discretizes. The learned units were modeled to focus on the “most important” factors to represent the speech audio (Baevski et al., 2020). W2V2 features have already been shown to work on several speech tasks, such as phoneme recognition, speech emotion recognition, and mispronunciation detection (Baevski et al., 2020; Pepino et al., 2021; Xu et al., 2021). As with dysfluencies, the “most important” parts of speech are disturbed. We therefore hypothesize that these features are suitable for dysfluency detection. We use a model pre-trained on 960 hours of unlabeled speech from the LibriSpeech corpus (Panayotov et al., 2015). The model was subsequently fine-tuned for automatic speech recognition (ASR) on the transcripts of the same data. The weights of this model were published by the authors of W2V2 (Baevski et al., 2020).

We extract W2V2 vectors for each audio sample for our experiments, yielding a 768-dimensional feature vector for every 20ms of raw audio. The W2V2 models allow the extraction of features at different layers in the feature encoder. Each layer yields a different representation that might be more or less suitable for a task than a later or previous layer (Baevski et al., 2021). The selected extraction layer is a tunable hyperparameter in this setup. We then take the mean of all vectors per clip and use this as the input to train an SVM classifier with a Gaussian kernel.

### 4.3. LSTM and LSTM-Attention classifier

For our baseline LSTM-model for stutter detection we follow the baseline single target learning approach described in (Lea et al., 2021). The two networks described here consist of a single layer long short-term memory (LSTM) layer with a hidden size of 64 neurons. The last hidden state is fed into a fully connected layer for classification.

The LSTM-Attention classifier (LSTM-A) is an extension of this model. An attention module complements the model with one attention head (Vaswani et al., 2017). The attention module takes all hidden states of the LSTM module as inputs with respect to the last hidden state. The output of the attention module is then fed into the fully connected layer for classification instead of the last hidden state of the LSTM module. Both networks were trained with a single weighted cross-entropy loss term, using class weights.

Both models were trained with a batch size of 64 and an initial learning rate of 0.001 and the Adam optimizer. Early stopping was employed based on cross-validation error.

For the transfer learning (TL) experiments, we used weights from models that were pre-trained on the SEP-28k dataset with the training parameters specified in the original paper by (Lea et al., 2021), who used batch size of 256 with an initial learning rate of 0.01 and the Adam optimizer. As input features to both models, we use 40-dimensional mel-filterbank energy features with a window size of 25 ms, a frame step of 10 ms, and frequency cut-offs at 0 and 8000 Hz.

## 5. Experiments

The main distinction between the classification methods employed is the feature or feature-encoding method used. OpenSMILE is based on handcrafted acoustic features that explicitly model prior knowledge of speech; W2V2 is a state-of-the-art neural feature-encoder that has learned speech units that capture the essence of speech. The LSTM-based models were used as a sequence encoder for the traditional, signal-processing-based spectral features as a reference to the baseline system from (Lea et al., 2021).

### 5.1. Evaluation

All experiments use 5-fold cross-validation. We split the folds by speaker so no samples from a speaker in the

training fold will appear in the test fold. We report the mean F1 score per dysfluency over the five folds and the standard deviation in brackets. All annotations represent a binary label, and respectively all models were trained as binary classification systems in a one-vs-all approach. We also included the results of a fictitious random classifier using class priors for reference purposes as a lower bound for experimental results in Table 3.

### 5.2. Results

OpenSMILE results indicate that low-level acoustic descriptors can capture the phonetically striking dysfluency types as well as modifications.

Even though the average of all W2V2 features vectors per clip was used for classification, thereby completely ignoring the sequential nature of the problem, the SVM utilizing W2V2 features performs best consistently over all experiments and dysfluency types. It underlines the capability of these transformer models to capture the intricacies of speech.

The LSTM model performs poorly, which might be due to the complexity of the tasks. The LSTM-A model performs consistently better than the LSTM model, which is to be expected. The attention mechanism helps the model focus on the more relevant parts of the sequence by emphasizing these inputs. Still, both models perform below expectations, which might be due to too little training data. The initial cross-language transfer-learning experiments increase performance over all dysfluency types but word repetitions for both the LSTM and LSTM-A system. These results support this assumption.

Modifications are the class that is consistently detected best regardless of model. This is not surprising as it is a very distinctive pattern and a learned behaviour shared among all PWS recorded for this dataset, probably making it the most straightforward pattern to detect. Performance of all approaches is consistently worst detecting word repetition. One factor is the small number of positive examples, only 3.8 % of clips in the corpus, which makes training difficult. Another factor is that of all the types of stuttering, word repetitions need the longest context to be recognizable. At the same time, they are acoustically almost indistinguishable from non-dysfluent speech, as they are just a repetition of a word. The pattern is complex as it spans large parts of a clip. It could probably be detected most reliably with an ASR system and an adapted language model, as suggested by (Alharbi et al., 2017).

## 6. Discussion and Outlook

We presented KSoF, a new resource of stuttered speech including speech of PWS who learned to modify their speech. This unique resource will enable more research into the automatic assessment of stuttering severity in a therapeutic context. The rather simple labeling approach proves to be a reliable and cost-effective by uti-

Features	System	Mod	Bl	Int	Pro	Snd	Wd
-	<b>Random</b>	0.096	0.071	0.029	0.0258	0.038	0.003
<b>openSMILE</b>	<b>SVM</b>	0.58 (0.20)	0.40 (0.14)	0.34 (0.07)	0.32 (0.09)	0.36 (0.10)	0.05 (0.07)
<b>wav2vec 2.0</b>	<b>SVM</b>	<b>0.73</b> (0.05)	<b>0.57</b> (0.11)	<b>0.59</b> (0.08)	<b>0.40</b> (0.03)	<b>0.43</b> (0.12)	<b>0.17</b> (0.04)
<b>Mel-Filterbank</b>	<b>LSTM</b>	0.36 (0.13)	0.25 (0.09)	0.23 (0.05)	0.19 (0.04)	0.22 (0.16)	0.10 (0.02)
	<b>LSTM (TL)</b>	0.42 (0.10)	0.32 (0.11)	0.25 (0.04)	0.22 (0.01)	0.23 (0.10)	0.10 (0.02)
	<b>LSTM-A</b>	0.52 (0.09)	0.39 (0.10)	0.30 (0.10)	0.26 (0.04)	0.16 (0.06)	0.10 (0.04)
	<b>LSTM-A (TL)</b>	0.53 (0.08)	0.45 (0.12)	0.37 (0.05)	0.29 (0.04)	0.26 (0.15)	0.10 (0.02)

Table 3: Classification results are reported in the format **mean (std)** per metric for each of the labels related to stuttering: Modified (Mod), blocks (Bl), interjections (Int), prolongations (Pro), sound repetitions (Snd), and word repetitions (Wd).

lizing naive annotators. It could be shown that even little training can help the inter-rater agreement and consequently also the reliability of labels.

Baseline machine learning experiments – while promising – raise questions to be addressed by future research. It is unclear why the spectral features and the LSTM system performed below expectations, especially when comparing results to the hand-crafted heuristic features extracted with openSMILE. A detailed error analysis per feature and dysfluency type could shed light on this. Prolongations are the dysfluency type that is phonetically most similar to modified speech, which might be a factor for the rather poor baseline performance. A detailed analysis of misclassified clips can help to answer these questions. A detailed look at how W2V2 encodes the audio might also reveal why the predictive power of those features is so big, even when averaging values over whole clips.

For future work, we plan to explore multi-class classification of stuttering with a single classifier for all types of stuttering. Future research should focus on detecting and localizing of dysfluency events in continuous speech. This can help to improve ASR systems by identifying people with dysfluent speech, and speech therapy applications can profit from precise automated feedback. We also encourage researchers to explore different aspects of the data, such as the recording situation that might lead to alternative experimental settings.

### Acknowledgment

We thank Joëlle Döring for providing the initial annotation and segmentation and the rewarding discussions of all aspects of fluency and dysfluency. This work was supported by the Bayerisches Wissenschaftsforum (BayWISS).

## 7. Bibliographical References

Alharbi, S., Simons, A. J. H., Brumfitt, S., and Green, P. (2017). Automatic recognition of children’s read speech for stuttering application. In *6th Workshop on Child Computer Interaction (WOCCI 2017)*, pages 1–6. ISCA, November.

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A Framework for Self-

Supervised Learning of Speech Representations. In H. Larochelle, et al., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Baevski, A., Hsu, W.-N., Conneau, A., and Auli, M. (2021). Unsupervised Speech Recognition. *arXiv:2105.11084 [cs, eess]*, October. arXiv: 2105.11084.

Bayerl, S. P., Hönig, F., Reister, J., and Riedhammer, K. (2020). Towards Automated Assessment of Stuttering and Stuttering Therapy. In *Text, Speech, and Dialogue*, volume 12284. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

Bernstein Ratner, N. and MacWhinney, B. (2018). Fluency Bank: A new resource for fluency research and practice. *Journal of Fluency Disorders*, 56:69–80, June.

Cordes, A. K. and Ingham, R. J. (1994). Time-Interval Measurement of Stuttering: Effects of Interval Duration. *Journal of Speech, Language, and Hearing Research*, 37(4), August.

Ellis, J. B. and Ramig, P. R. (2009). A Handbook on Stuttering. *Journal of Fluency Disorders*, 34(4):295–299, December.

Euler, H. A., Gudenberg, A. W. v., Jung, K., and Neumann, K. (2009). Computergestützte Therapie bei Redeflussstörungen: Die langfristige Wirksamkeit der Kasseler Stottertherapie (KST). *Sprache · Stimme · Gehör*, 33(04):193–202, December.

Howell, P., Davis, S., and Bartrip, J. (2009). The University College London Archive of Stuttered Speech (UCLASS). *Journal of Speech, Language, and Hearing Research*, (2).

Ingham, R. J., Ingham, J. C., Bothe, A. K., Wang, Y., and Kilgo, M. (2015). Efficacy of the Modifying Phonation Intervals (MPI) Stuttering Treatment Program With Adults Who Stutter. *American Journal of Speech-Language Pathology*, 24(2):256–271, May.

Karimi, H., O’Brian, S., Onslow, M., and Jones, M. (2014). Absolute and Relative Reliability of Percentage of Syllables Stuttered and Severity Rating Scales. *Journal of Speech, Language, and Hearing Research*, 57(4):1284–1295, August.

- Kourkounakis, T., Hajavi, A., and Etemad, A. (2020). 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)*, pages 6089–6093. IEEE.
- Kourkounakis, T., Hajavi, A., and Etemad, A. (2021). FluentNet: End-to-End Detection of Stuttered Speech Disfluencies With Deep Learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2986–2999.
- Landis, J. R. and Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159, March.
- Lea, C., Mitra, V., Joshi, A., Kajarekar, S., and Bigham, J. P. (2021). 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)*, pages 6798–6802, Toronto, ON, Canada, June. IEEE.
- Lickley, R. (2017). Disfluency in typical and stuttered speech. *Fattori sociali e biologici nella variazione fonetica*, (3):373.
- Mongia, M., Gupta, A., Vijay, A., and Sadhu, R. (2019). Management of stuttering using cognitive behavior therapy and mindfulness meditation. *Industrial Psychiatry Journal*, 28(1):4.
- Noeth, E., Niemann, H., Haderlein, T., Decher, M., Eysholdt, U., Rosanowski, F., and Wittenberg, T. (2000). Automatic stuttering recognition using hidden Markov models. In *Sixth International Conference on Spoken Language Processing*, volume 4, pages 65–68, January.
- Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). 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, April. IEEE.
- Pepino, L., Riera, P., and Ferrer, L. (2021). Emotion Recognition from Speech Using wav2vec 2.0 Embeddings. In *Interspeech 2021*, pages 3400–3404. ISCA, August.
- Miquel S. Porta, et al., editors. (2014). *A dictionary of epidemiology*. Oxford University Press, Oxford, six edition edition.
- Riad, R., Bachoud-Lévi, A.-C., Rudzicz, F., and Dupoux, E. (2020). Identification of Primary and Collateral Tracks in Stuttered Speech. In *Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, France, May. European Language Resources Association.
- Schuller, B., Steidl, S., Batliner, A., Hirschberg, J., Burgoon, J. K., Baird, A., Elkins, A., Zhang, Y., Coutinho, E., and Evanini, K. (2016). The INTER-SPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity & Native Language. In *Proc. Interspeech 2016*.
- Schuller, B. W. \., Batliner, A., Bergler, C., Mascolo, C., Han, J., Lefter, I., Kaya, H., Amiriparian, S., Baird, A., Stappen, L., Ottl, S., Gerczuk, M., Tzirakis, P., Brown, C., Chauhan, J., Grammenos, A., Hasthanasombat, A., Spathis, D., Xia, T., Cicuta, P., Rothkrantz, M \, L. J. \., Zwerts, J., Treep, J., and Kaandorp, C. (2021). The INTER-SPEECH 2021 Computational Paralinguistics Challenge: COVID-19 Cough, COVID-19 Speech, Escalation & Primates. In *Proceedings INTERSPEECH 2021, 22nd Annual Conference of the International Speech Communication Association*, Brno, Czechia, September. ISCA.
- Valente, A. R. S., Jesus, L. M. T., Hall, A., and Leahy, M. (2015). Event- and interval-based measurement of stuttering: a review: Event- and interval-based measurement of stuttering: a review. *International Journal of Language & Communication Disorders*, 50(1):14–30, January.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is All you Need. In I. Guyon, et al., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Xu, X., Kang, Y., Cao, S., Lin, B., and Ma, L. (2021). Explore wav2vec 2.0 for Mispronunciation Detection. In *Interspeech 2021*, pages 4428–4432. ISCA, August.
- Yairi, E. and Ambrose, N. (2013). Epidemiology of stuttering: 21st century advances. *Journal of Fluency Disorders*, 38(2):66–87, June.
- Świetlicka, I., Kuniszyk-Józkowiak, W., and Smółka, E. (2013). Hierarchical ANN system for stuttering identification. *Computer Speech & Language*, 27(1):228–242, January.