KSoF: The Kassel State of Fluency Dataset

A Therapy Centered Dataset of Stuttering

Sebastian P. Bayerl[1], Alexander Wolff von Gudenberg[3], Florian Hönig[3], Elmar Nöth[2], and Korbinian Riedhammer[1]

[1]Technische Hochschule Nürnberg Georg Simon Ohm
[2]Friedrich-Alexander Universität Erlangen-Nürnberg
[3]Institut der Kasseler Stottertherapie

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. 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. This work describes the creation of a unique resource containing stutteren speech and modified speech.

Dataset Profile

- The audio was recorded at the *Institut der Kasseler Stottertherapie*.
- Dataset containing speech from persons who stutter (PWS)
- 37 speaker (28 male, 9 female)
- 5597 3-sec long clips labeled and annotated with:
  - six stuttering related labels (blocks, prolongations, sound repetitions, word repetitions, interjections, and speech modifications)
  - six meta labels
  - metadata (recording situation, therapy status, microphone type, task, speaker, gender)

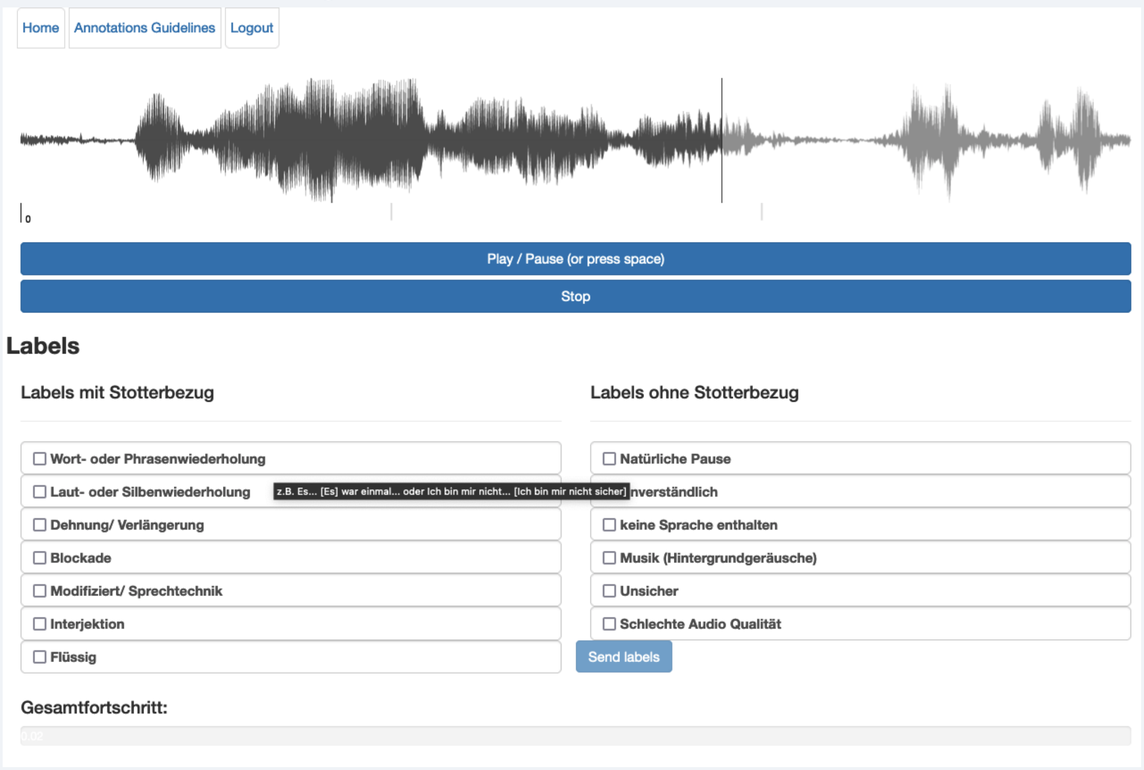
Labeling Process

- Two step process with three annotators, who were naive listeners at the time
- Multiple binary choice
- Six types of stuttering related labels
- Six meta labels
- hard task
- one hour agreement meeting substantially improved inter-rater reliability

Table 1: Fleiss' kappa agreement statistics for each type of stuttering.

Stuttering Labels	KSoF	Test	SEP-28k
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	-

Figure 1: Screenshot of the annotation tool



Contact

arXiv

Kassel State of Fluency

Table 2: Distribution of annotations in the Kassel State of Fluency (KSoF) and SEP-28k for reference

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, e.g., "[IIII]I", otherwi[ssss]se
Sound Repetition	14.76 %	8.3 %	Repeated syllables, e.g., "[nat-nat-nat-]naturally", or sounds; "I [t-t-t-]talked to dad.
Word/ Phrase Rep.	3.88 %	9.8 %	"I have [I have] done no such thing"
No dysfluencies	24.75 %	56.9 %	There are no audible dysfluencies
Modified	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", "also"

Baseline Experiments

- openSMILE features as a common baseline method
- wav2vec 2.0 embeddings extracted from ASR fine tuned model
  - SVM as classifier
  - point of extraction influences results
- Single layer LSTM classifiers used for comparison to SEP-28k
  - Mel-Filterbank spectrograms (40-dim, 25ms window, 10ms frame step)
  - Attention mechanism added
  - Cross lingual transfer learning

Table 3: Classification results are reported in the format **mean (std)**

System	Mod	Bl	Int	Pro	Snd	Wd
Random	0.096	0.071	0.029	0.0258	0.038	0.003
SVM + oS	0.58 (0.20)	0.40 (0.14)	0.34 (0.07)	0.32 (0.09)	0.36 (0.10)	0.05 (0.07)
SVM + w2v2	**0.73** (0.05)	**0.57** (0.11)	**0.59** (0.08)	**0.40** (0.03)	**0.43** (0.12)	**0.17** (0.04)
LSTM	0.36 (0.13)	0.25 (0.09)	0.23 (0.05)	0.19 (0.04)	0.22 (0.16)	0.10 (0.02)
LSTM (TL)	0.42 (0.10)	0.32 (0.11)	0.25 (0.04)	0.22 (0.01)	0.23 (0.10)	0.10 (0.02)
LSTM-A	0.52 (0.09)	0.39 (0.10)	0.30 (0.10)	0.26 (0.04)	0.16 (0.06)	0.10 (0.04)
LSTM-A (TL)	0.53 (0.08)	0.45 (0.12)	0.37 (0.05)	0.29 (0.04)	0.26 (0.15)	0.10 (0.02)

- Consistently best results using wav2vec 2.0 features
- results for modifications are consistently best
  - Pattern is very distinct
  - acquired speech pattern, universally taught, therefore less variance expected
- Word repetitions consistently worst
  - number of labeled instances in KSoF might be the reason
  - needs largest temporal context, acoustically indistinguishable from normal speech
- performance on prolongations might suffer because of similarity to modifications

Summary

- We created a unique resource containing stuttering therapy data, including labels and metadata
- minimal training can improve inter-rater-reliability of naive annotators, enabling a cost effective labeling option
- Baseline experiments using state of the art classification systems still leaves room for improvement

References

[1] Colin Lea, Vikramjit Mitra, Aparna Joshi, Sachin Kajarekar, and Jeffrey 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, June 2021.

[2] Sebastian P Bayerl, Florian Hönig, Joëlle Reister, and Korbinian Riedhammer. Towards automated assessment of stuttering and stuttering therapy. In *International Conference on Text, Speech, and Dialogue*. Springer, Cham, 2020.