Automatic Assessment of Signs of Alcohol Dependency Syndrome from Spontaneous Speech

Gábor Gosztolya[1,2], András Bence Lázár[3],
Ildikó Hoffmann[3,4], Otília Bagi[3],
Fruzsina Fanni Farkas[3], Janka Gardics[3],
László Tóth[2], János Kálmán[3]

[1]HUN-REN--SZTE Research Group on AI, Szeged, Hungary
[2]University of Szeged, Institute of Informatics, Szeged, Hungary
[3]University of Szeged, Department of Psychiatry, Szeged, Hungary
[4]HUN-REN Research Cener for Linguistics, Budapest, Hungary

The effect of Alcohol

- Alcohol is a progressive central nervous system depressant
- Alcohol dependence can affect executive functions
- The motor and cognitive functions might be affected as well…
- …along with impairing executive functions, affecting speech production:
  - Verbal fluency
  - Working memory
  - Recent memory
  - Visuospatial abilities
  - Visual recognition and processing speed

Contribution of this Study

- Short-term influence of alcohol is widely studied
  - i.e. is the speaker drunk?
- Long-term effects are rarely investigated
- Alcohol Dependency Syndrome (ADS)
  - We focus the long-term effects of alcohol consumption on speech
- In this study we
  - Present a speech corpus with 35 ADS speakers and 35 healthy controls, having two spontaneous speech tasks
  - We automatically distinguish the two speaker groups by machine learning
  - We also distinguish the recordings of the two speech tasks
  - We investigate the extent of pauses present in the speech of the subjects

Speech Recordings

- Subjects
  - 35 ADS, 35 healthy controls (HC), Hungarian native speakers
  - No statistically significant differences in age, gender & education
- Two separate speech tasks
  - As a neutral topic, describe the events of their **previous day**
  - As an **alcohol-related** speech task, describe their relationship to alcohol and situations where they found it hard to resist drinking
- The duration of the recordings is the following (in sec):

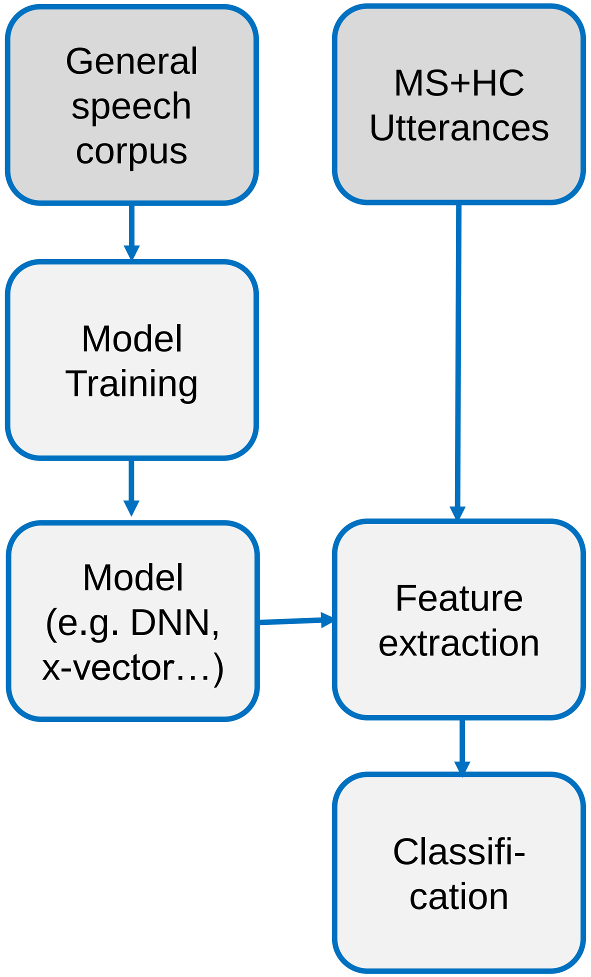
Speech task	ADS	HC
Previous day	76.7 ± 45.0	84.1 ± 33.3
Alcohol-related	80.9 ± 45.8	86.4 ± 31.4

Automatic Speech Analysis of ADS subjects

- Due to data scarcity, end-to-end models are difficult to use
  - E.g. 70 subjects in our case (3h 11m total duration)
  - For cross-validation (nested cross-validation) we have to train lots of (DNN) models
  - In general, this is the case in the pathological speech processing area
- Due to this, feature extraction and classification are typically distinct steps
- We focus on “general” (i.e. not task-specific) features
  - Like i-vectors, x-vectors, ECAPA-TDNN…
  - Standard approach for detecting Parkinson’s or Alzheimer’s Disease, depression, etc.

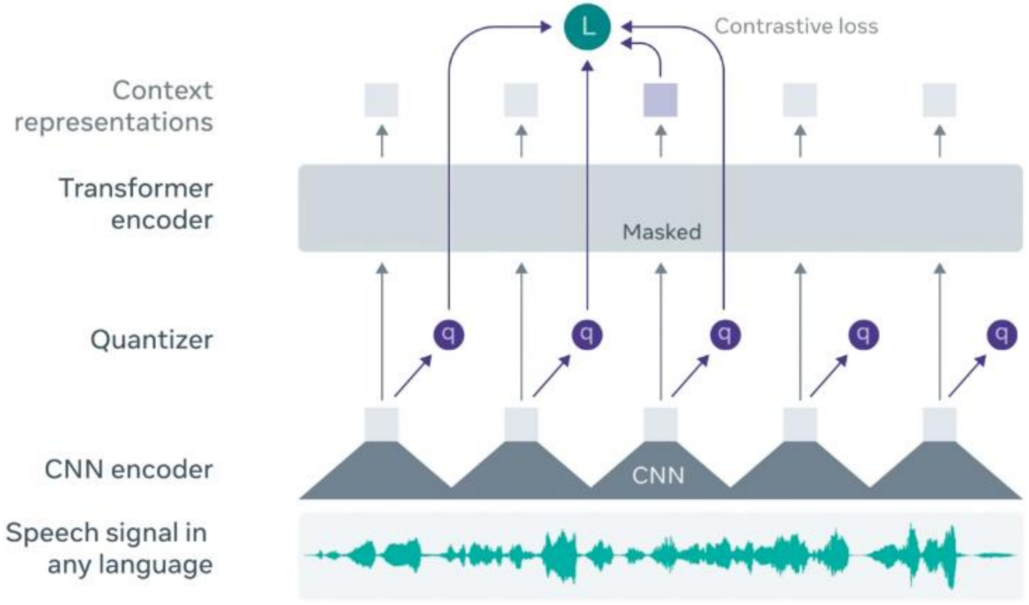
Automatic Speech Analysis of ADS subjects

- Even feature extraction is split into two steps
  - 1) we train (or fine-tune) some model (on a general corpus)
  - 2) by using the model on the actual utterances, we extract (model-specific) features from them
- From another point of view
  - 1) we build a “general” model for “normal speech”
  - 2) we express (with the features) the difference of the given utterance and this “normal speech”



wav2vec 2.0

- CNN encoder
  – Converts the raw speech signal into a latent representation
- Transformer encoder
  – Transformer layer, its output is the contextualized representation
- Linear projection layer
  – Obtained by fine-tuning for the final task (e.g. ASR for the given language)
- Cross-lingual Representation Learning (XLSR) wav2vec 2.0
  – For tasks with limited **unlabeled** data: we pre-train the model for multiple languages simultaneously



Experimental Setup

- wav2vec 2.0 model
  - wav2vec2-large-xlsr53-hungarian (from Huggingface)
  - Fine-tuned on the Hungarian part of the Common Voice 6.1 corpus (8 hrs)
- Feature extraction: embeddings from last hidden layers of blocks
  - Convolutional & contextualized (“fine-tuned”)
  - Frame-level embeddings → mean, standard deviation
  - 512, 1024 → 1024 (convolutional) and 2048 (fine-tuned) features
- Classification: SVM
  - libSVM, linear kernel, 35-fold **nested** cross-validation, repeated 5 times
- Evaluation: EER (Equal Error Rate), AUC (Area under ROC)
- Significance tests: Mann-Whitney U test

Results (ADS vs. HC)

Speech task	Embedding	EER	AUC
Previous day	Convolutional	11.4%	0.947
	Fine-tuned	20.0%	0.885
Alcohol-related	Convolutional	16.6%	0.906
	Fine-tuned	9.1%	0.982

- The results are overall quite good
  - Probably ADS changes the subjects' speech, which can be detected
- The speech tasks are similarly useful
  - Convolutional embeddings work better for the Previous day task ($p < 0.01$)
  - Fine-tuned embeddings work better for the Alcohol-related task ($p < 0.01$)
- Overall, the results for the Alcohol-related task are a bit better

Results (Previous day vs. Alcohol-related)

Subjects	Embedding	EER	AUC
ADS	Convolutional	39.4%	0.605
	Fine-tuned	43.4%	0.576
HC	Convolutional	16.6%	0.892
	Fine-tuned	14.9%	0.893
ADS + HC	Convolutional	31.7%	0.699
	Fine-tuned	22.3%	0.829

- **ADS** subjects: the results are barely better than random
  - Convolutional embeddings were slightly better
  - EER: $p = 0.0397$, AUC: $p > 0.05$
  - Probably there was not a huge difference in the speech during the two speech tasks (or it was not captured by the wav2vec 2.0 embeddings)

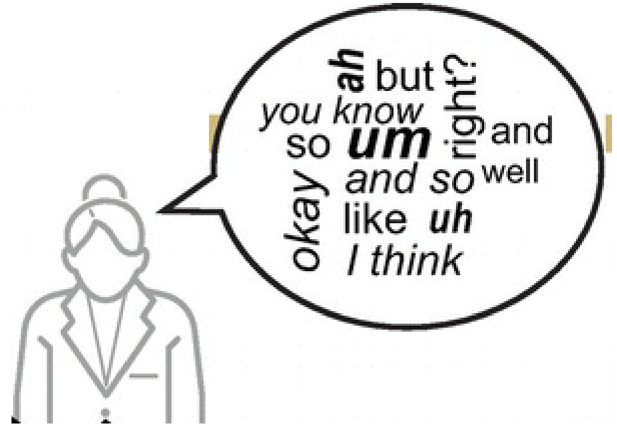
Results (Previous day vs. Alcohol-related)

Subjects	Embedding	EER	AUC
ADS	Convolutional	39.4%	0.605
	Fine-tuned	43.4%	0.576
HC	Convolutional	16.6%	0.892
	Fine-tuned	14.9%	0.893
ADS + HC	Convolutional	31.7%	0.699
	Fine-tuned	22.3%	0.829

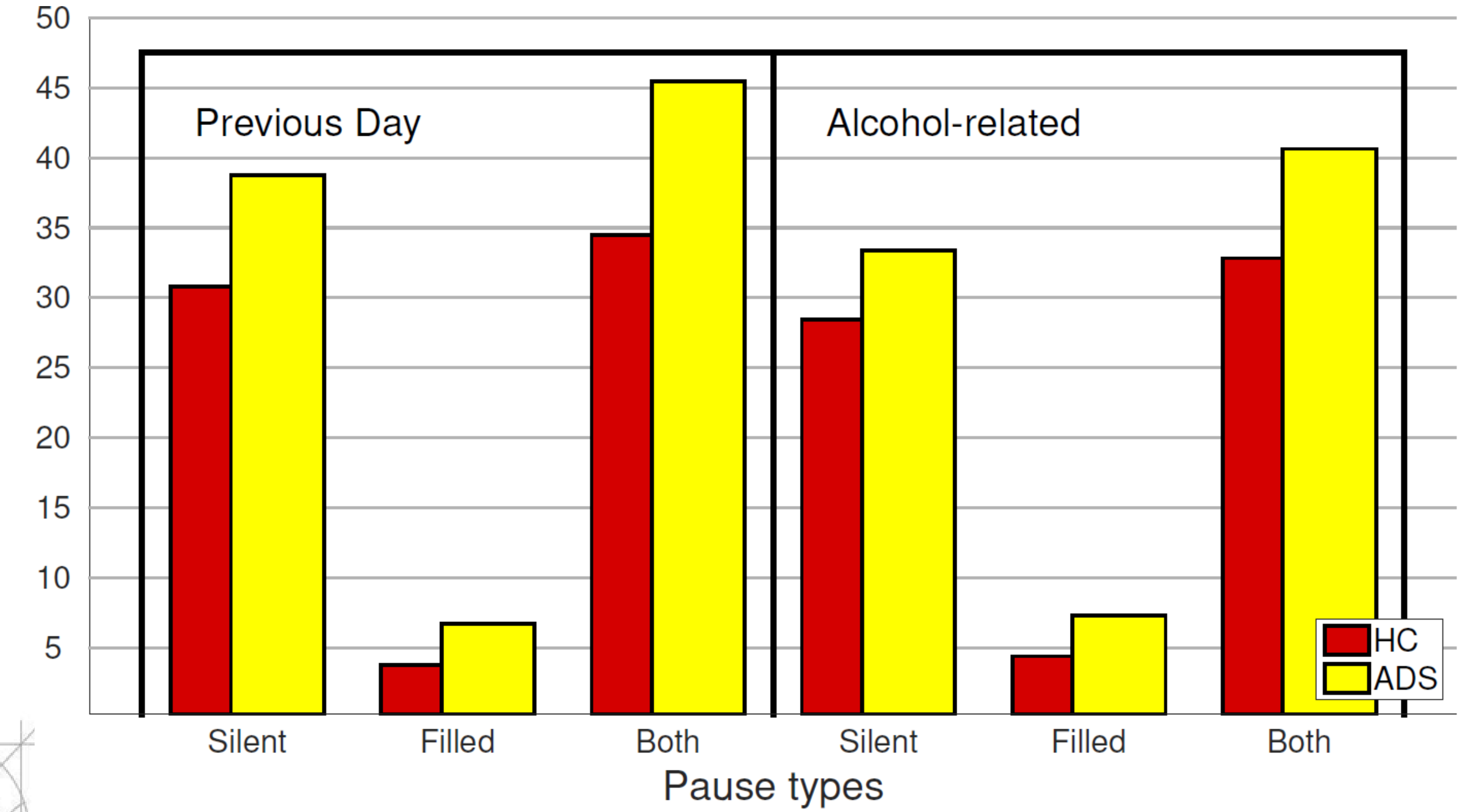
- **HC** subjects: the results are overall quite good
  - Both with convolutional and fine-tuned embeddings ($p > 0.05$)
- **ADS + HC** subjects: the results are in-between
  - Fine-tuned embeddings were significantly better ($p < 0.01$ for EER & AUC)

Investigating the Amount of Pauses

- Lastly, we investigated usefulness of the amount of pauses
  - Silent pauses and filled pauses (“er”, “um” etc.), with durations ≥ 30 ms
- Calculated by a standard HMM/DNN hybrid model
  - The acoustic model was trained on 60 hours of Hungarian spontaneous speech (increased to 240 hours by noise augmentation)
  - Phone-level speech recognition
  - Filled pause was treated as a special phone
  - Amount of duration (%) was calculated over the whole utterance



Amount of Pauses Produced



- ADS subjects, in general, produced more pauses than healthy controls
- This is true for all three pause types (“silent”, “filled”, “both”) and all speech tasks

- ADS subjects also produced more silent pauses in the Previous day speech task than in the Alcohol-related speech task

Classification Results with Pause Stats

Classification task	Data	EER	AUC
ADS vs. HC	Previous day	21.1%	0.826
	Alcohol-related	33.1%	0.730
Previous day vs. Alcohol-related	ADS	57.4%	0.409
	HC	53.4%	0.431
	ADS + HC	55.0%	0.497

- Classification experiments with only the three pause statistics as features
- Experimental setup is the same

- The two speaker groups could efficiently be separated
- The two speech tasks were indistinguishable
  - EER > 50%, AUC < 0.5
  - On the figure, the two speech tasks had similar pause characteristics
  - However, the ADS subjects clearly produced more silent pauses

Summary

- We presented a speech corpus with 35 ADS and 35 HC subjects
  - Speech tasks: a neutral topic (previous day) and an alcohol-related one
- We tried to automatically distinguish the two speaker groups
  - A standard workflow: wav2vec 2.0 embeddings + SVM, cross-validation
- We tried to distinguish the two speech tasks
  - They proved to be quite similar for the ADS speakers, but quite different for the HC subjects
- We measured the amount of pauses
  - Silent and filled pauses, detected by a HMM/DNN hybrid model
  - Besides a manual investigation of the tendencies, we also performed classification experiments

Limitations

- The number of subjects (35 + 35) is not that high
  – Although it is a common-sized corpus for pathological speech processing
- The wav2vec 2.0 model was fine-tuned on a limited amount of data (only 8 hours)
- Only the last hidden layers of the two wav2vec 2.0 blocks (convolutional and contextualized) were used
- Further **interpretable** attributes? (Just like the amount of pauses)

This study was supported by the NRDI Office of Hungary (grants TKP2021-NVA-09 and RRF-2.3.1-21-2022-00004), and by the Géza Hetényi Grant (SZTE-ÁOK-KKA-2024-HG)