Investigating the clusters discovered by pre-trained AV-HuBERT

Aalto University
School of Electrical Engineering
USSEE
Understanding speech and scene with ears and eyes

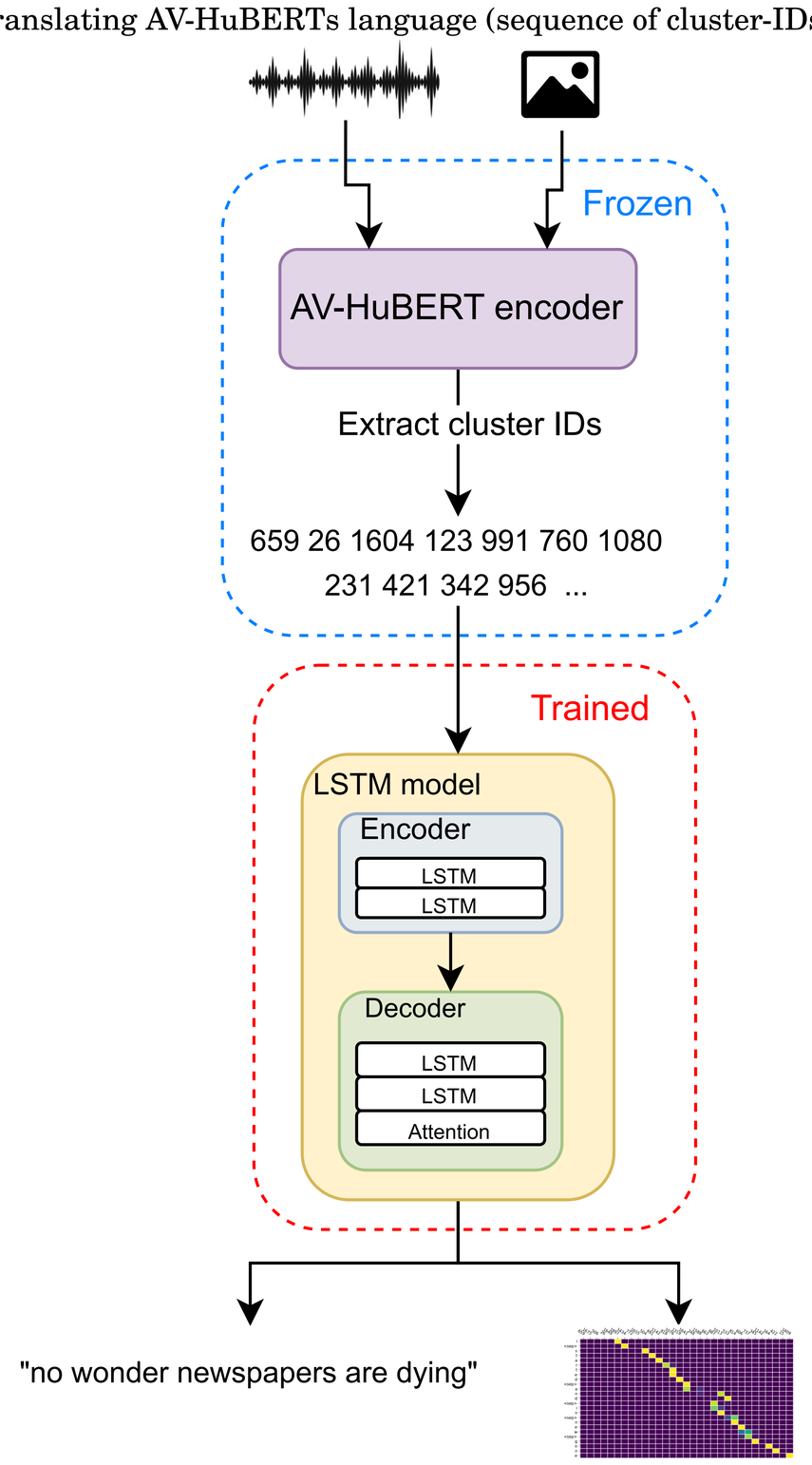
Anja Virkkunen, Marek Sarvaš, Guangpu Huang, Tamas Grosz, Mikko Kurimo
Department of Information and Communications Engineering, Aalto University, Finland

Motivation

- Pre-trained AV-HuBERT has demonstrated excellent results
- The latent clusters discovered during un-supervised training are usually ignored and discarded by the fine-tuning algorithm
- Hypothesis: latent clusters contain valuable information for AV-ASR
- Using only the clusters could lead to low-resource solutions
- Linking latent clusters and characters improves transparency

Proposed solution

Translating AV-HuBERTs language (sequence of cluster-IDs) to English



- Training data inferred only once with AV-HuBERT.
- Consecutive duplicate IDs were removed.
- Attention maps were extracted for further analysis.

Experimental details

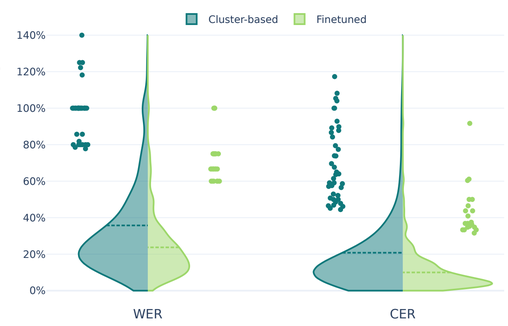
- For our experiment, the LRS3-TED data was used (30h of videos)
- Both *base* and *large* models were investigated, after the 4th and 5th pre-training phase
- Models pre-trained on 407h of videos
- Finetuned models' decoder part consists of 6 Transformer layers
- **Fairseq** used for fine-tuning, BPE targets, Adam optimizer, Label smooting + Cross Entropy
- **OpenNMT** for translation training and inference
- LSTM-based model with 500 or 1024 units, character outputs
- Attention maps extracted on the test data and aggregated for analysis

Audiovisual ASR results

Model	Size	WER (%)
Finetuned base iter 4	161M	**7.35**
Finetuned base iter 5	161M	7.75
Finetuned large iter 5	477M	8.84
base iter 4 + LSTM500	10M	23.65
base iter 4 + LSTM1024	42M	24.56
base iter 5 + LSTM500	11M	18.92
base iter 5 + LSTM1024	43M	20.97
large iter 5 + LSTM500	11M	**17.85**
large iter 5 + LSTM1024	43M	19.15
k-means base iter 5 + LSTM500	11M	**17.55**
k-means base iter 5 + LSTM1024	43M	20.93
k-means large iter 5 + LSTM500	11M	35.80
k-means large iter 5 + LSTM1024	43M	39.77

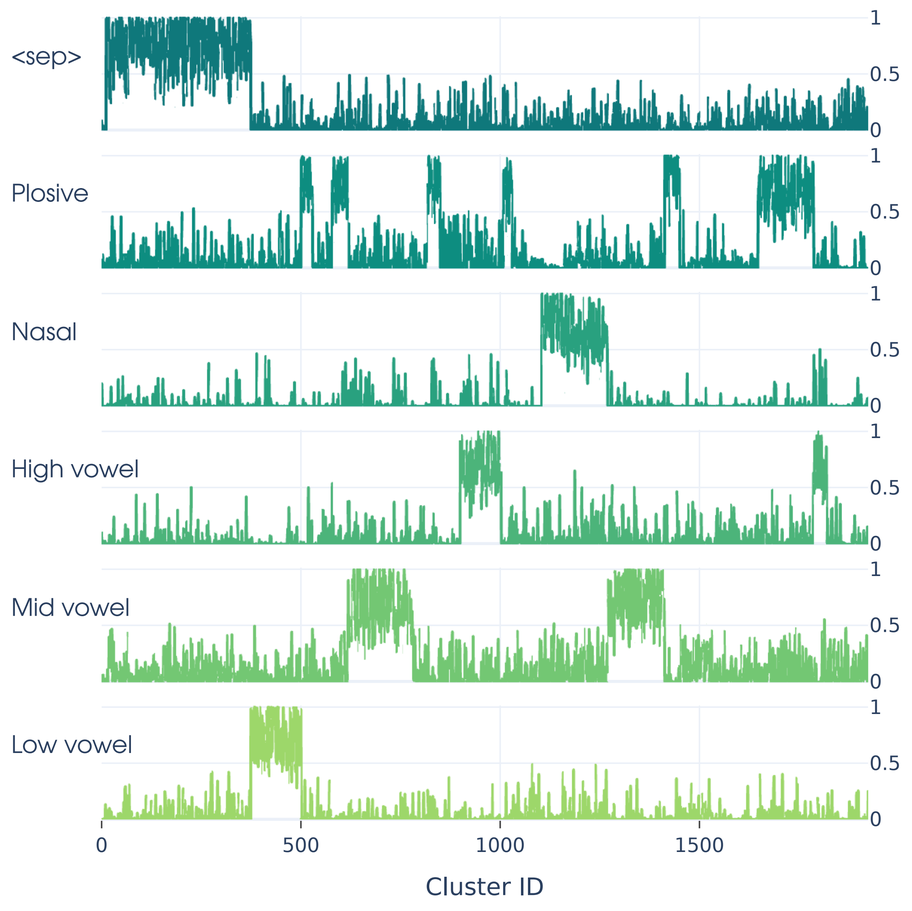
Hallucinations

~5% of the test data is affected by severe hallucinations (insertion of word sequences)
Main source: common pronouns (*we, they*), conjunction (*or*) and the preposition *of*, causing repetition of short phrases



What are the clusters used for?

Using the attention values, we can match clusters to characters.



Observations

- *E*, *M* and *T* had the most latent groups, 10 unused clusters
- There was a clear <*bos*> and multiple <*eos*> clusters

Conclusions and Future Work

- Translating latent clusters into text offers a low-resource AV-ASR solution
- Latent clusters could be linked to characters/phonetic groups to improve transparency, but hallucinations are a problem
- Many clusters are dedicated to other non-ASR tasks, what are they used for?