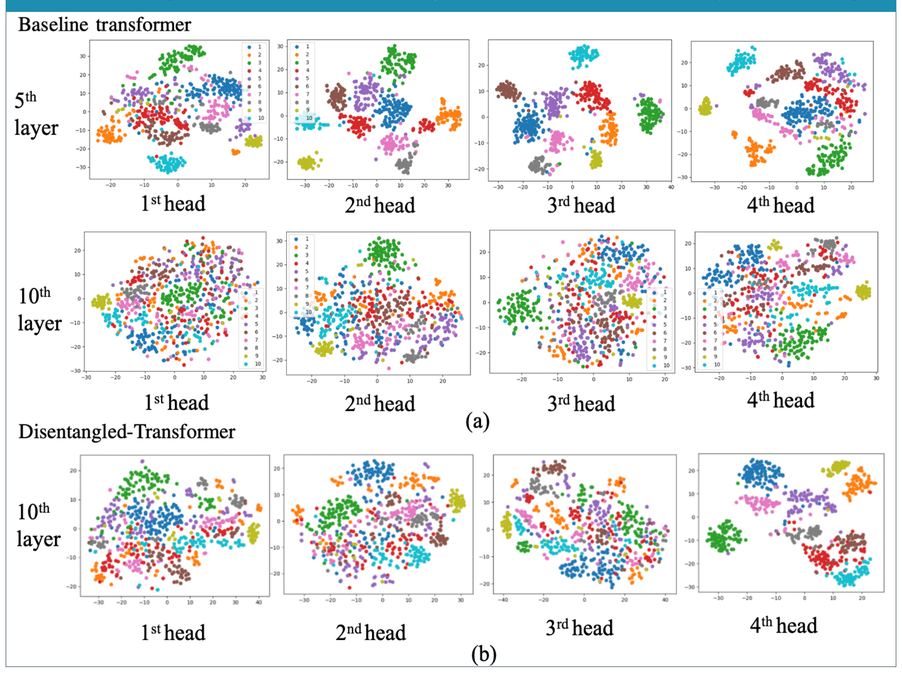
Disentangled-Transformer: An Explainable End-to-End Automatic Speech Recognition Model with Speech Content-Context Separation Learning Based on Varying Temporal Resolutions

Pu Wang, Hugo Van hamme
KU Leuven, Dept. of Electrical Engineering, PSI, Leuven, Belgium

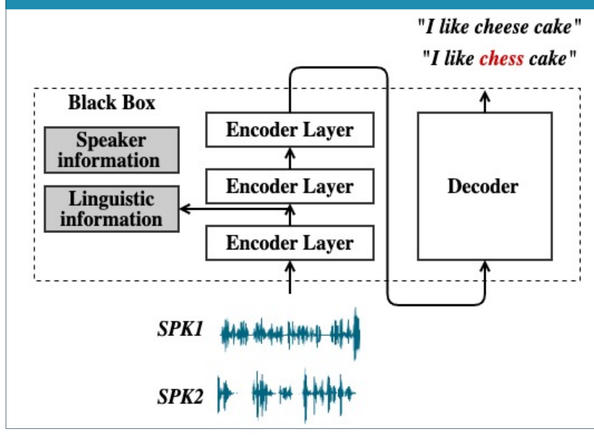
Motivation

- Deep learning methods are highly **data-driven**;
- **Data mismatches** (e.g. healthy speakers versus dysarthric speakers) **lead to** automatic speech recognition (**ASR**) **errors**;
- **Linguistic information or speech content is expected to remain consistent across different speakers;**
- Speech models often capture **entangled speech traits** (speech content, speaker, accent, …) in the learned representations at each encoder layer. Fig (a) shows **arbitrary** and **inconsistent entanglement** can occur at various components or layers within the entangled model;
- Linguistic information is **entangled** with speaker information and **changes** due to speaker mismatch;
- **Lack of interpretability** when errors occur.

T-SNE (each color represents one speaker)



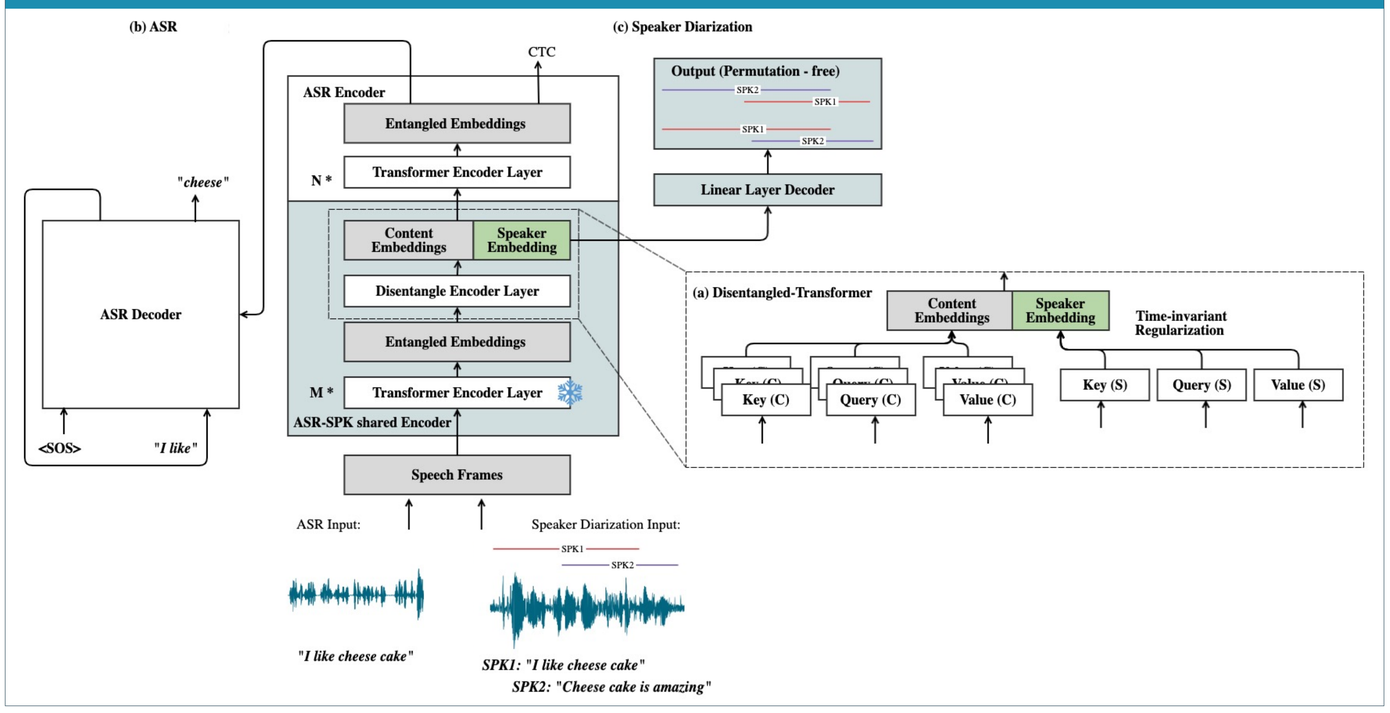
Entangled ASR model



Aim

- To answer: "**Which** aspects of the learned representations contribute to fluctuations in recognition performance, and **where** in the network architecture do they originate?" -> **enhancing the explainability of learned representations**;
- **Disentangle the internal representation** of the ASR encoder **into sub-embeddings**, with each sub-embedding explicitly correlated with a specific speech trait (e.g. speech content and speaker identity) **without compromising** ASR performance.

Disentangled-Transformer



Method

- Useful information changes at different rates: **speech content** (what) - **rapid** change, **speaker** (who) – **slower** pace;
- Disentangle speech trait-specific sub-embeddings based on the different **temporal behavior**;
- **Disentangled-Transformer** with **explicit sub-embedding heads**: Introduces **time-invariant regularization** to penalize rapid changes in the speaker-embedding attention head of the **specific encoder layer(s)** during ASR training;
- Share the Disentangled encoder with an end-to-end (E2E) speaker diarization task.

ASR (LibriSpeech 100h) & Speaker Diarization (LibriMix)

- Disentangled-Transformer introduces regularization to all 18 encoder layers;

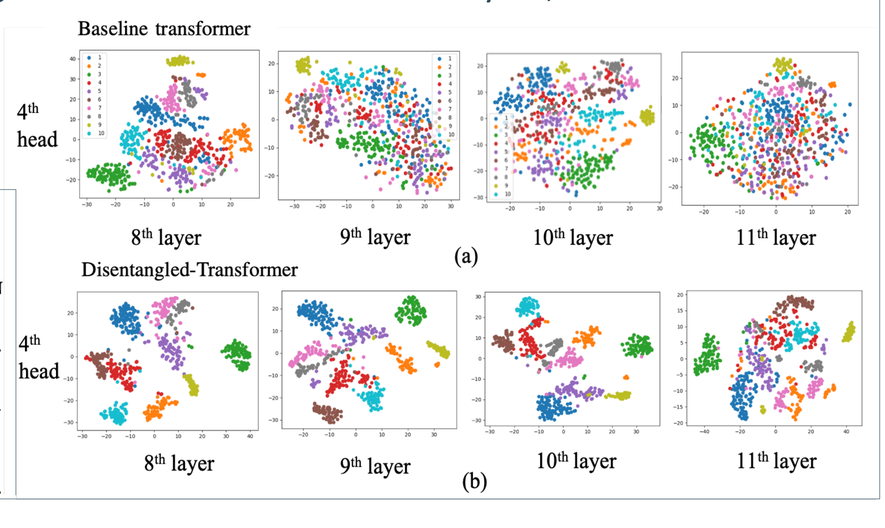
WER (%) FOR THE BASELINE TRANSFORMER AND DISENTANGLED-TRANSFORMER, TRAINED ON LIBRISPEECH100H

Method	dev_clean	dev_other	test_clean	test_other
Baseline transformer	8.0	20.1	8.3	20.6
Disentangled-Transformer	7.8	19.6	8.1	20.0

- Regularization is applied to the 4th layer.

DER (%) FOR THE BENCHMARK TRANSFORMER, DISENTANGLED-TRANSFORMER, AND ASR-TRANSFORMER TRAINED ON LIBRIMIX OVER 5 AND 10 EPOCHS

Method	5 epochs		10 epochs	
	dev_clean	test_clean	dev_clean	test_clean
Benchmark Transformer	19.2	19.4	9.8	9.0
Disentangled-Transformer	10.7	10.6	5.6	5.6
ASR-transformer	18.5	21.6	13.3	12.9



- Disentanglement does not degrade the ASR performance;
- Disentangled-Transformer explicitly shows clear speaker patterns in the constrained heads and layers (Fig (b));
- Disentangled speaker embeddings improve the diarization performance.

Acknowledgments

The research was supported by the Flemish Government under "Onderzoeksprogramma AI Vlaanderen" and FWO-SBO grant S004923N: NELF.