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

Pu Wang[0000−0001−8725−6225] and Hugo Van hamme[0000−0003−1331−5186]

KU Leuven, Department of Electrical Engineering-EAST,
Kasteelpark Arenberg 10, 3001, Belgium
{pu.wang, hugo.vanhamme}@esat.kuleuven.be

Abstract. Modern automatic speech recognition (ASR) systems, leveraging transformer-based end-to-end (E2E) data-driven deep learning methods, have markedly enhanced recognition accuracy, paving the way for advanced speech recognition applications such as Apple Siri. However, deep learning applications are often perceived as black boxes. For example, in E2E ASR models, the learned representations at each encoder layer often capture multiple speech traits that are highly entangled - not only the intended speech content (the actual linguistic information) but also the speaker identity, dialect, accent, emotion, background noise, and other contextual factors. As a result, the learned representations of current E2E ASR models lack interpretability.

In the real world, speech data encompass significant diversity, including variations in speaker and speaking environments. [1] shows that an ASR model trained on one group of speakers but tested on different groups exhibits significant fluctuations in recognition performance. While the linguistic information or speech content is expected to remain consistent across different speakers, the model produces speech content-related representations that are entangled with other speech traits not present in the training data, such as speaker identity. Furthermore, our experimental results show that with a well-trained transformer-based E2E ASR model, the representations extracted from certain attention heads of the first several encoder layers form clear speaker and gender clusters, while other layers or attention heads do not exhibit the same behavior. This arbitrary and inconsistent entanglement can occur at various components or layers within the model, making it challenging to answer the question, "*Which* aspects of the learned representations from *where* in the network architecture contribute to fluctuations in recognition performance?". Consequently, this entanglement makes it difficult to improve the ASR model's generalization.

Our studies focus on enhancing the explainability of learning representations in transformer-based E2E ASR models. Specifically, we disentangle the internal representation of the encoder into sub-embeddings with each sub-embedding explicitly correlated with a specific speech trait (such as speaker identity) based on the different temporal behavior of each speech trait. For example, the *what* (linguistic information) is a *rapidly* varing embedding that requires a resolution of a few tens of milliseconds, while the *who* (speaker, accent, dialect) varies at a *slower* rate. In this study, we mainly discuss disentangling the content and speaker traits, a model we name Disentangle-Transformer. We introduce time-invariant regularization to penalize rapid changes in the one attention head of the Disentangle-Transformer's encoder layer during training. This regularization can be applied to either a single layer or the full set of encoder layers.

Experiments with training on LibriSpeech 100 dataset show that the Disentangle-Transformer, with an explicitly entangled speaker trait, can achieve the same performance as the vanilla transformer model, with a minor improvement observed. The word error rate (WER) drops from 8.0 to 7.8 on the dev-clean dataset, from 20.1 to 19.6 on dev-other, from 8.3 to 8.1 on test-clean, and from 20.6 to 20.0 on test-other. T-SNE plots of representations extracted from each attention head of each encoder layer of the Disentangle-Transformer and the vanilla transformer show that the Disentangle-Transformer consistently exhibits clear speaker patterns in the constrained attention head and layer, while the vanilla transformer does not show regular patterns, as some layers and attention heads contain more speaker traits.

To assess the extent of the explainability of the Distentangle-Transformer ASR model, we use the encoder of the trained Disentangle-Transformer with a single-layer linear decoder for speaker diarization on the LibriMix dataset, which overlaps utterances from two speakers in the LibriSpeech 100 dataset. Zero-shot speaker diarization using representations extracted from the time-invariant constrained attention head, referred to as the *disentangled speaker trait*, achieves a diraization error rate (DER) of around 22%, while representations extracted from the non-constrained attention head, referred to as the *disentangled content trait*, show a DER higher than 40%. With further fine-tuning of only the constrained layer of the Distentangle-Transformer on the speaker diarization target for 5 epochs, the disentangled speaker trait achieves a DER of around 11%. After training for 13 epochs (three hours), it reached a DER of around 7%, matching the performance of the benchmark model, which was trained on LibriMix dataset for 100 epochs.

Keywords: Explainable AI · Speech representation disentanglement · Speech recognition · Speaker diarization.

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

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Wang, P., Van Hamme, H.: Benefits of pre-trained mono-and cross-lingual speech representations for spoken language understanding of dutch dysarthric speech. EURASIP Journal on Audio, Speech, and Music Processing **2023**(1), 15 (2023)