Poster Number: 4022/SPE-52.3

Towards End-to-End Integration of Dialog History for Improved Spoken Language Understanding

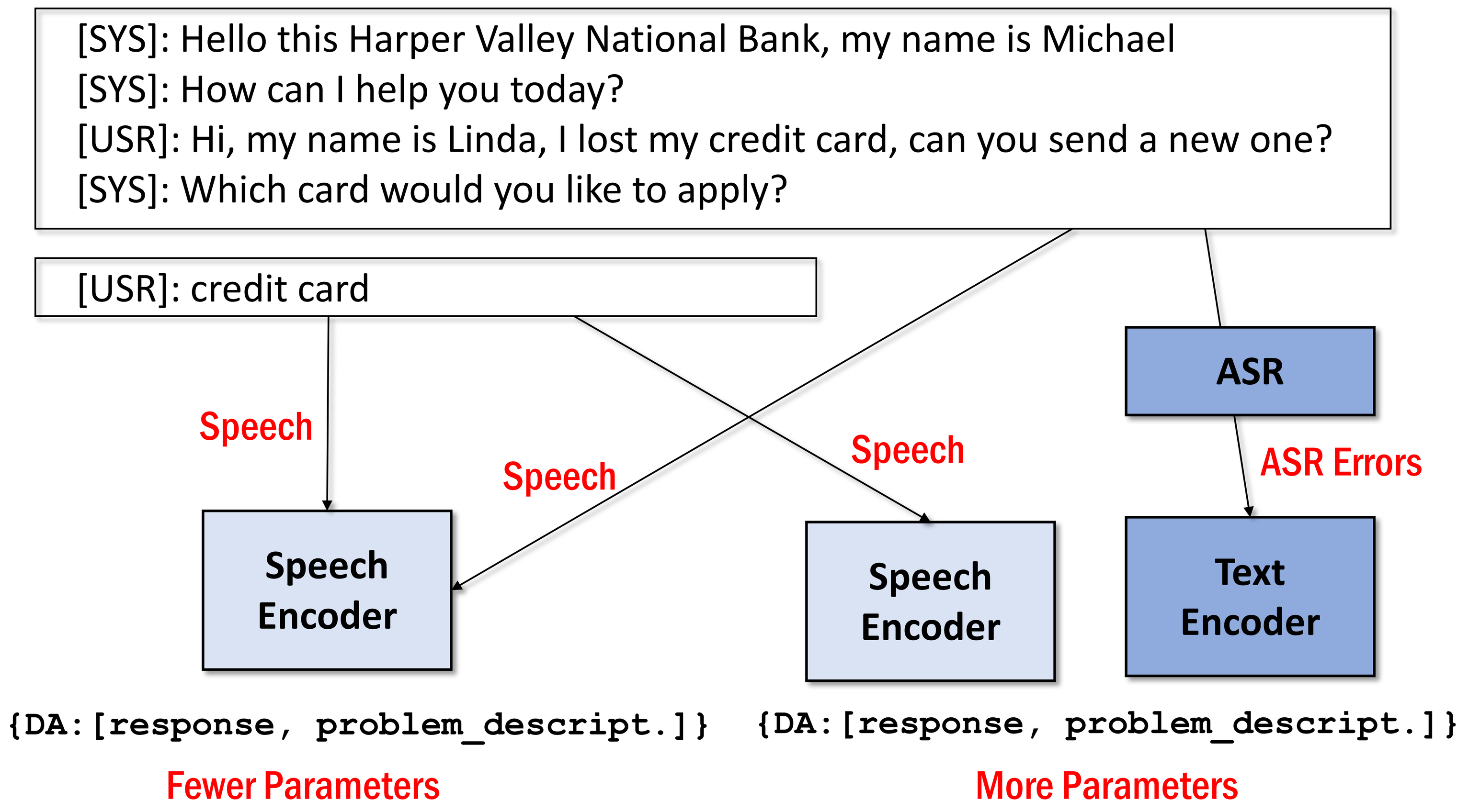
Vishal Sunder[1], Samuel Thomas[2], Hong-Kwang J. Kuo[2], Jatin Ganhotra[2], Brian Kingsbury[2], Eric Fosler-Lussier[1]

[1] The Ohio State University, [2] IBM Research

icassp 2022 Singapore

Motivation

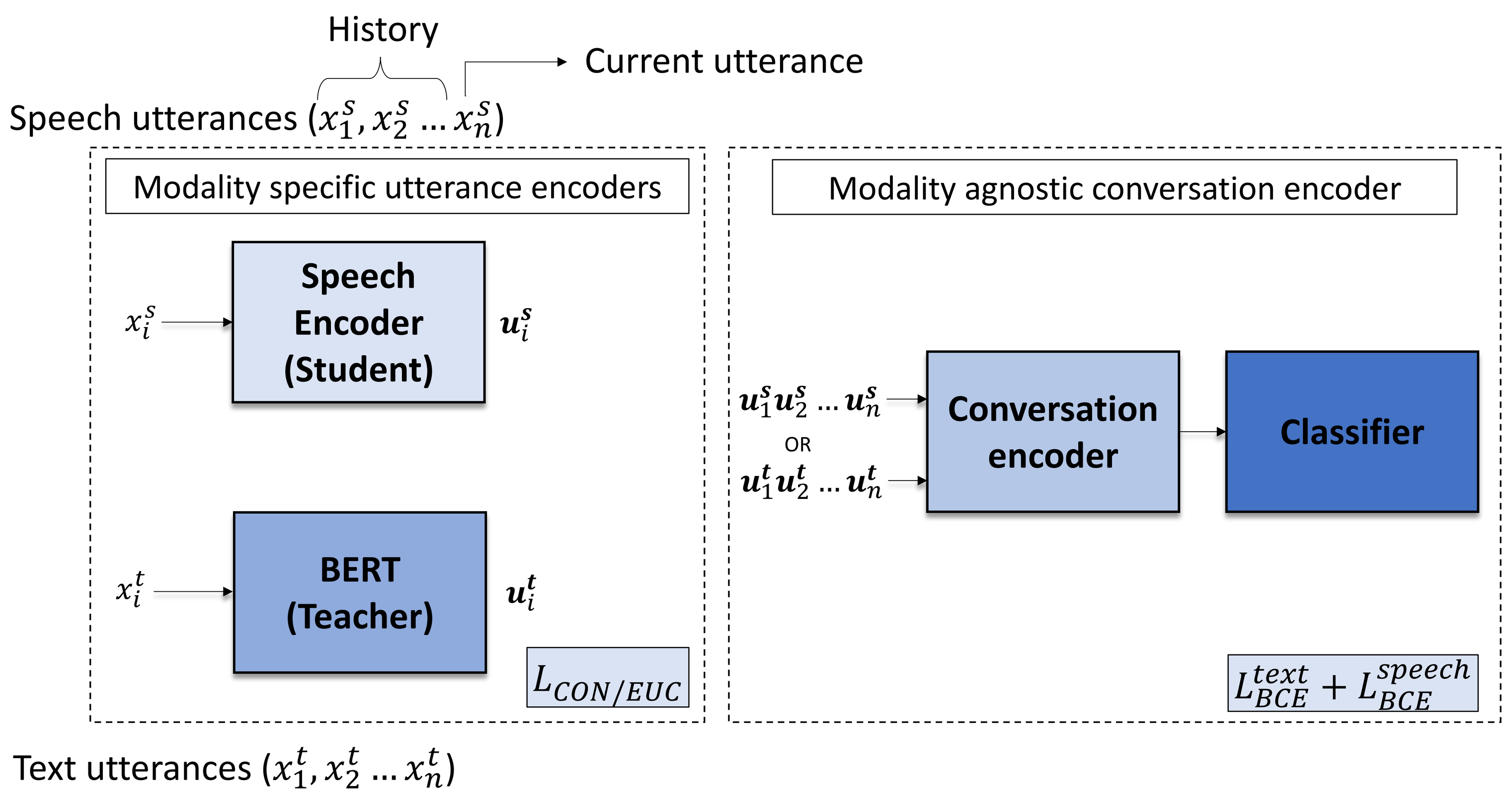
- Dialog history provides conversation context which is useful for dialog act classification in a spoken dialog system.
- E2E dialog systems have used dialog history in text form which needs a cascaded ASR [1]. This leads to an increase in model size.
- We propose a hierarchical model to integrate dialog history in speech form directly leading to significant improvements with 48% fewer parameters.



- We investigate the multi-label dialog act classification task on the HarperValleyBank dialog dataset.

Hierarchical Conversation Model

- The hierarchical model comprises of a low-level utterance encoder and a high-level conversation encoder which is modelled as a transformer.
- The utterance encoder is can be speech-based or text-based while the conversation encoder is independent of the modality.



- We can train 3 different types of model using the above architecture:
  1. **HIER-ST**: A model co-trained using both speech and text.
  2. **HIER-S**: A model trained using only speech.
  3. **HIER-T**: A model trained using only text.
- Semantic knowledge from text-based BERT is transferred to the speech-based utterance encoder using Euclidean loss (L_{EUC}) and Contrastive loss (L_{CON}).

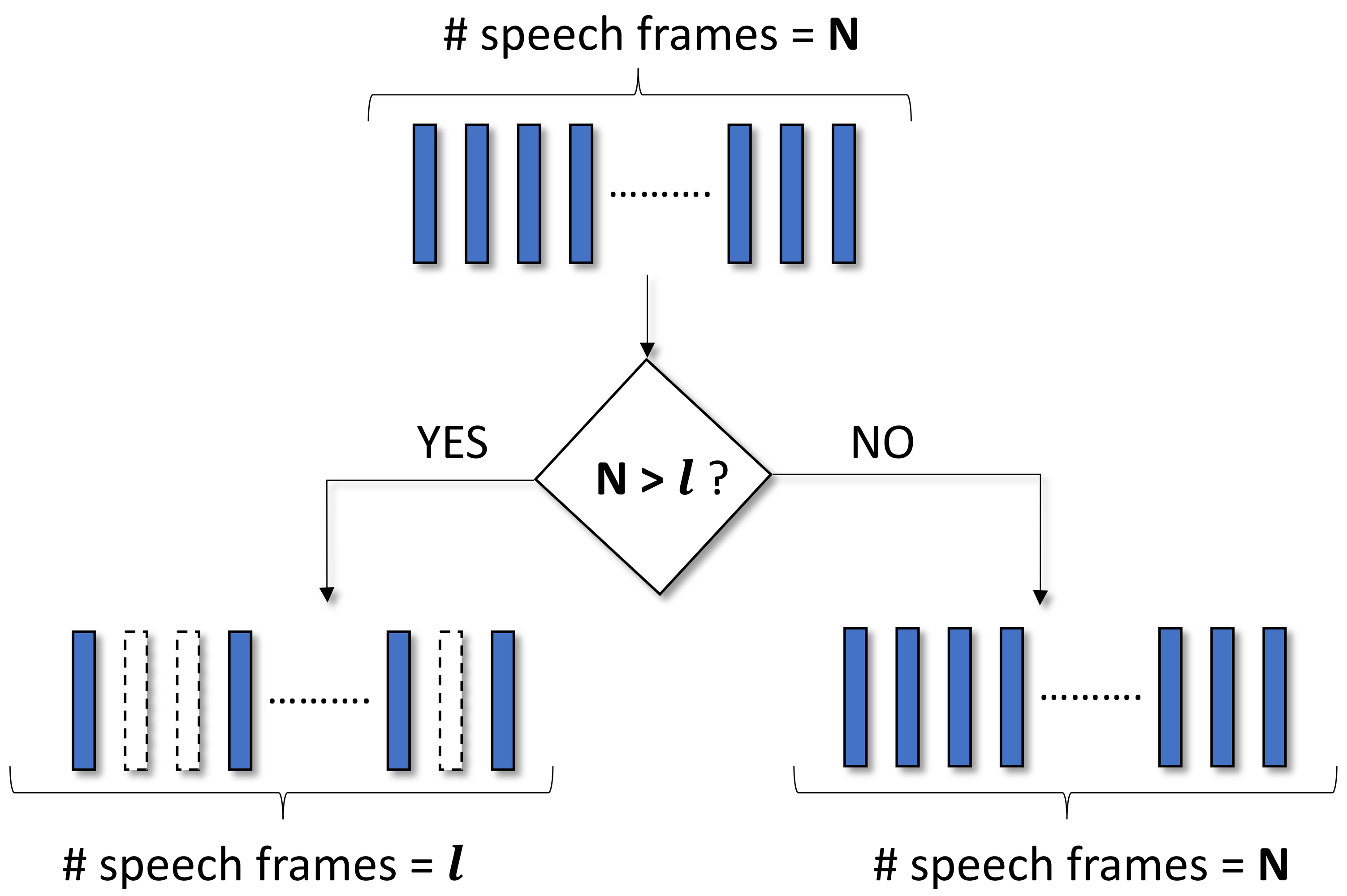
$$L_{EUC} = \frac{1}{|B|}\sum_{i=1}^{|B|} \left\| \mathbf{u}_N^s[i] - \mathbf{u}_N^t[i] \right\|_2$$

$$L_{CON} = -\frac{1}{2|B|}\sum_{i=1}^{|B|}\left(\log\frac{\exp(s_{ii})}{\sum_{j=1}^{|B|}\exp(s_{ij})} + \log\frac{\exp(s_{ii})}{\sum_{j=1}^{|B|}\exp(s_{ji})}\right)$$

where, s_{ij} represents the cosine similarity between two utterances in a batch.

DropFrame

- When dialog history is used in speech form, the speech sequence length can become very long which results in an increased training time for E2E models.
- Drop out random frames from a sequence of length $> l$.



# Frames (l)	Macro-F1	Train time
64	56.5	4
256	61.7	8
1024	60.2	25
All	59.9	27

- The training time is reduced significantly when frames are dropped. Also, performance is improved which shows that DropFrame also acts as an effective regularizer.

Experiments and Results

Experiments are done in two different settings.

Gold transcripts are available

- We use the speech-encoder (transcription network) from a fine tuned RNN-T ASR at the lower level [2].
- WER - 1.9%
- Gold transcripts are used for co-training HIER-ST.

Model	Macro-F1	# Params
(1T) BERT (on utterance)	56.1	168M
(2T) BERT (on context)	**63.5**	168M
(3T) HIER-T	63.3	200M
(1C) ASR $\longrightarrow$ BERT (on context)	**62.2**	168M
(2C) ASR $\longrightarrow$ HIER-T	61.3	200M
(1E) LSTM (on utterance)	54.0	54M
(2E) HIER-S	58.3	88M
(3E) HIER-ST	59.0	88M
(4E) HIER-ST + L_{EUC}	60.3	88M
(5E) HIER-ST + L_{CON}	**61.7**	88M
(6E) HIER-ST + L_{EUC} + L_{CON}	60.9	88M
(7E) HIER-ST + L_{CON} ($g(.;\phi)$ = LSTM)	61.3	**62M**

- **(5E)** *gives competitive performance compared to* **(1C)** *with significantly fewer parameters.*

Gold transcripts are not available

- We use the speech-encoder (transcription network) from an off-the-shelf RNN-T ASR at the lower level.
- WER - 11.3%
- ASR transcripts from the off-the-shelf ASR used for co-training HIER-ST.

Model	Macro-F1
(3C) ASR $\longrightarrow$ BERT (on context)	50.3
(8E) HIER-S	57.7
(9E) HIER-ST + L_{CON} (w/ ASR text)	**60.3**
(10E) HIER-ST + L_{CON} (w/ Gold text)	**61.7**

- HIER-ST *does not degrade in performance while the traditional cascaded model performs significantly worse due to acoustic mismatch.*

References

[1] J. Ganhotra, S. Thomas, H.-K. J. Kuo, S. Joshi, G. Saon, Z. Tüske, and B. Kingsbury, "Integrating dialog history into end-to-end spoken language understanding systems," *arXiv preprint arXiv:2108.08405*, 2021.

[2] G. Saon, Z. Tüske, D. Bolanos, and B. Kingsbury, "Advancing rnn transducer technology for speech recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5654–5658.