What next after ASR in Indian Languages? We speak in order to be understood!

Samuel Thomas
IBM Research AI

sthomas@us.ibm.com

MUCS 2021: MUltilingual and Code-Switching ASR Challenges for Low Resource Indian Languages

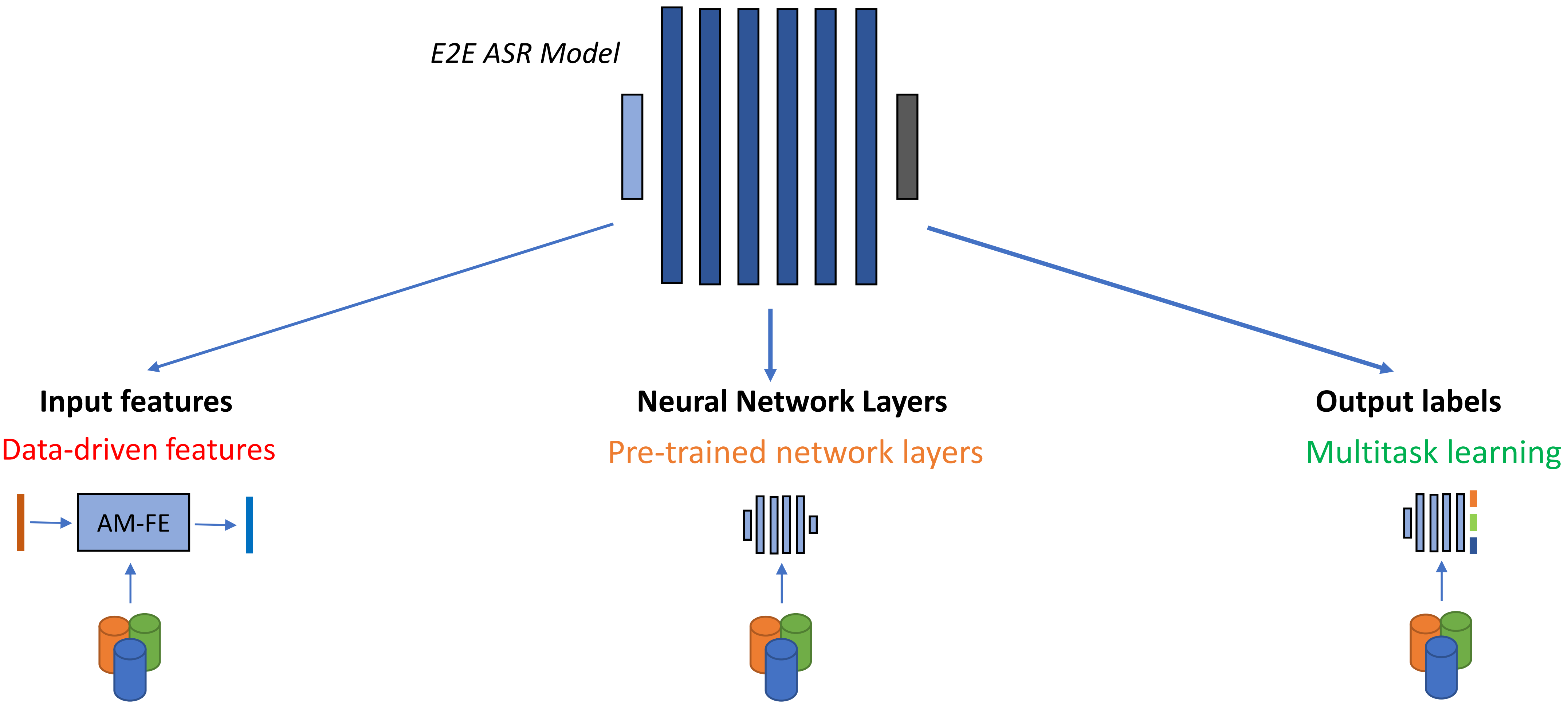
The Low-resource ASR Problem

Acoustic models for state-of-the-art speech recognition systems are typically trained on several hundred hours of task specific training data, or more. However, in low resource scenarios often only a few tens of hours of annotated training data are available.

How can we effectively build models in low resource settings?

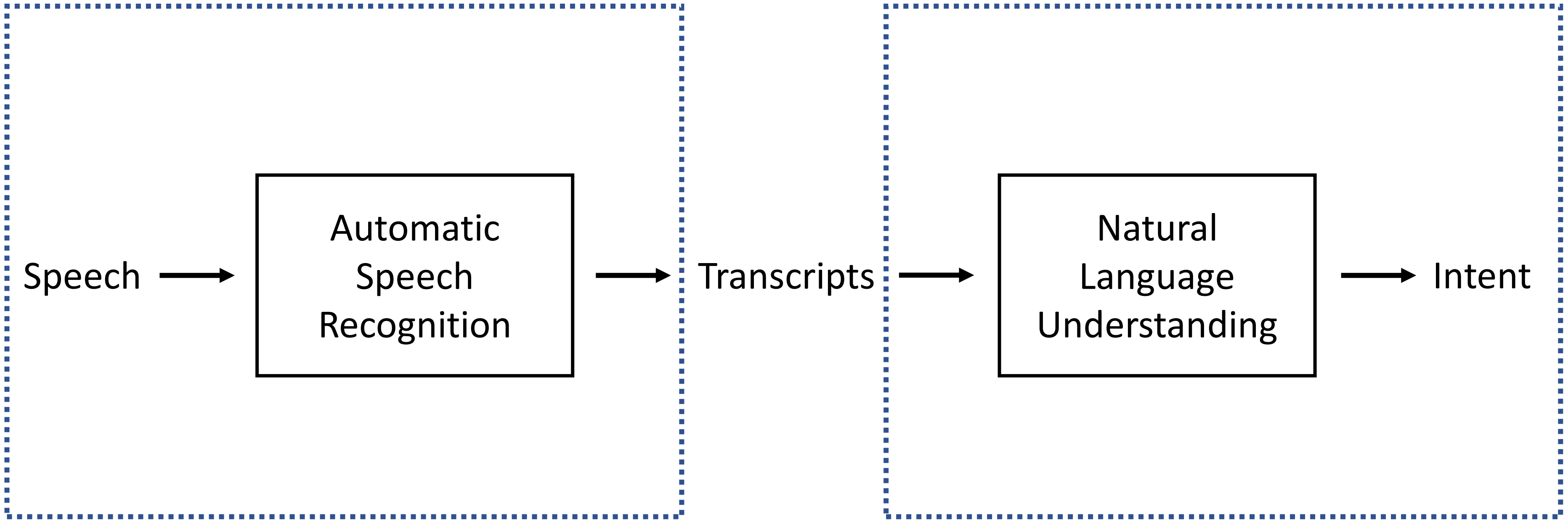
The Low-resource ASR Problem

How have we effectively built E2E ASR models in low resource settings?

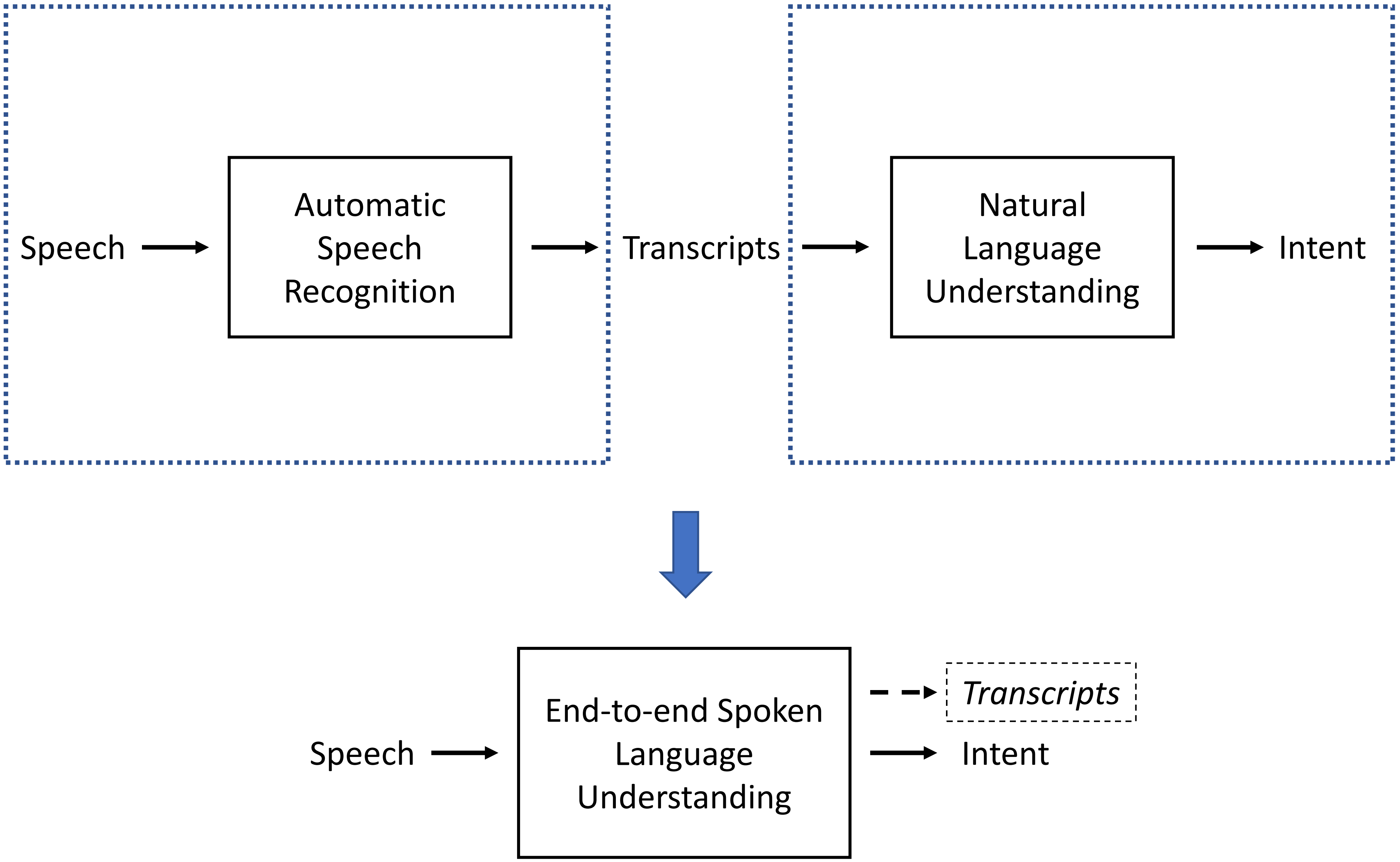


AM-FE – Acoustic model as a feature frontend

What do we do with ASR transcripts?



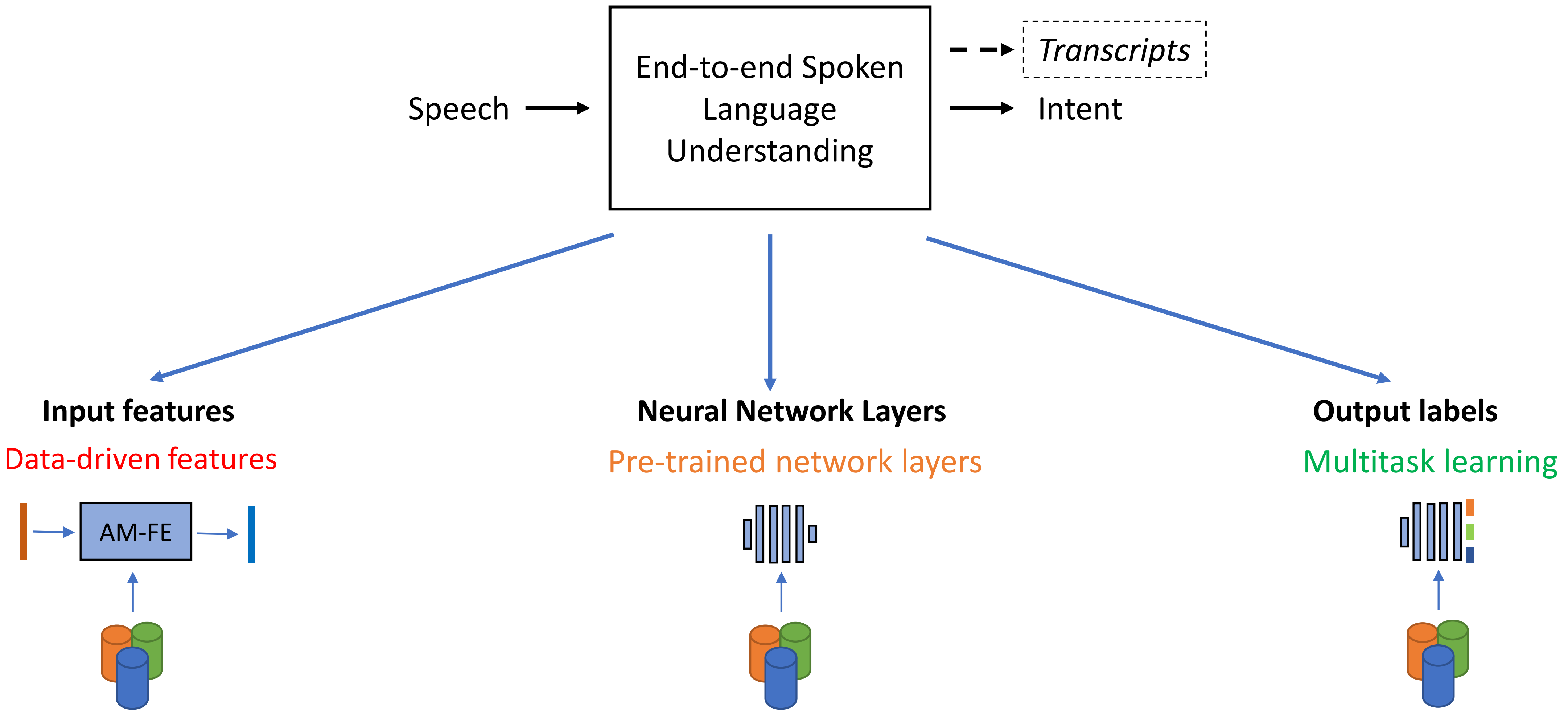
What next? End-to-end SLU



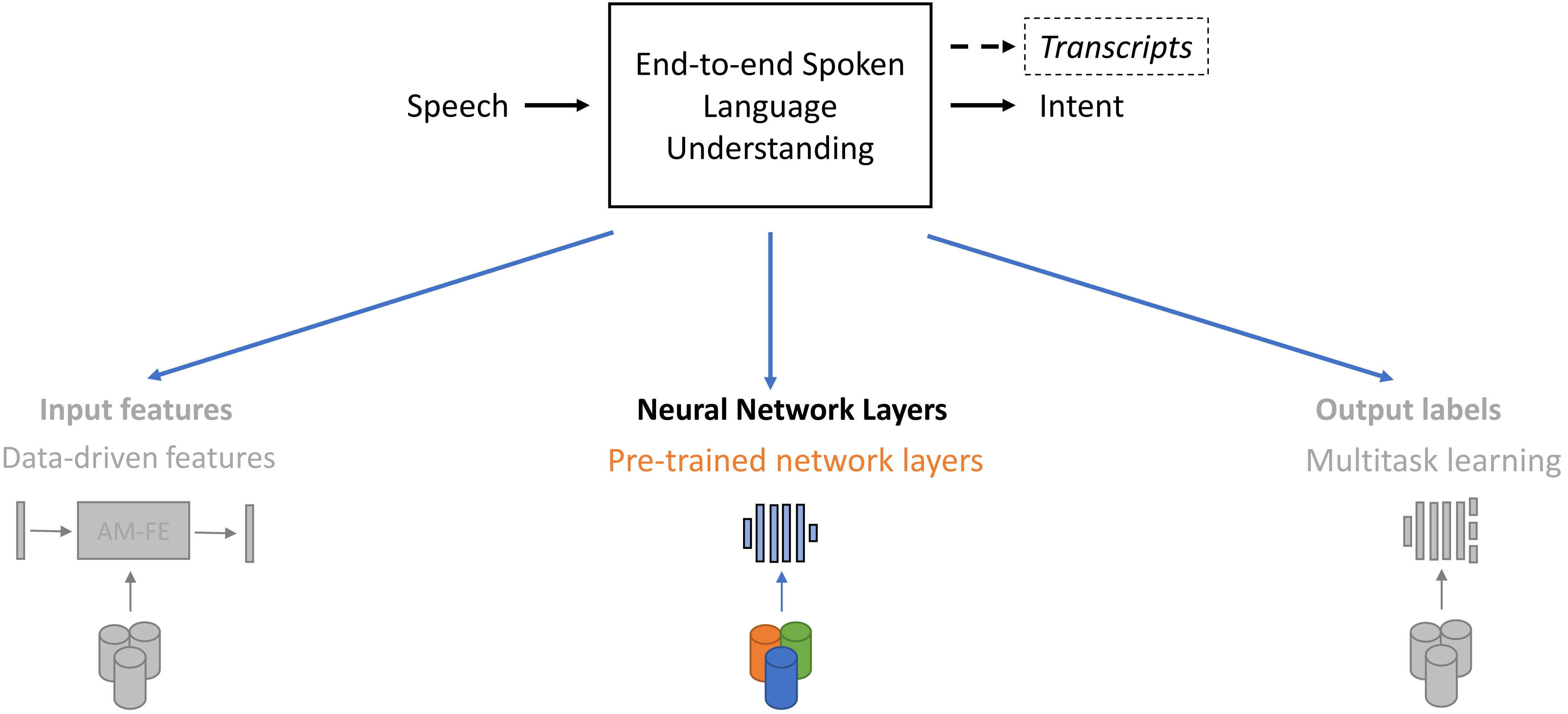
The SLU task as a low resource task

- Directly process speech to produce spoken language understanding (SLU) entity or intent label targets.
  - <speech> I want a flight to Delhi from Chennai that makes a stop in Mumbai
  - **<SLU> Transcript + Intent label**: I want a flight to Delhi from Chennai that makes a stop in Mumbai INT-FLIGHT
  - **<SLU> Transcript + Entity labels**: I want a flight to DELHI B-toloc.cityname from CHENNAI B-fromloc.cityname that makes a stop in MUMBAI B-stoploc.cityname
  - **<SLU> Entity labels only**: DELHI B-toloc.cityname CHENNAI B-fromloc.cityname MUMBAI B-stoploc.cityname
  - **<SLU> Intent label only**: INT-FLIGHT
- SLU as a low resource task
  - SLU domain specific data is typically very limited – few tens of hours
  - Data labelled with SLU intents and labels are also very limited.

Can we use what we learnt, for SLU?



Can we use what we learnt, for SLU?



Leveraging pre-trained networks

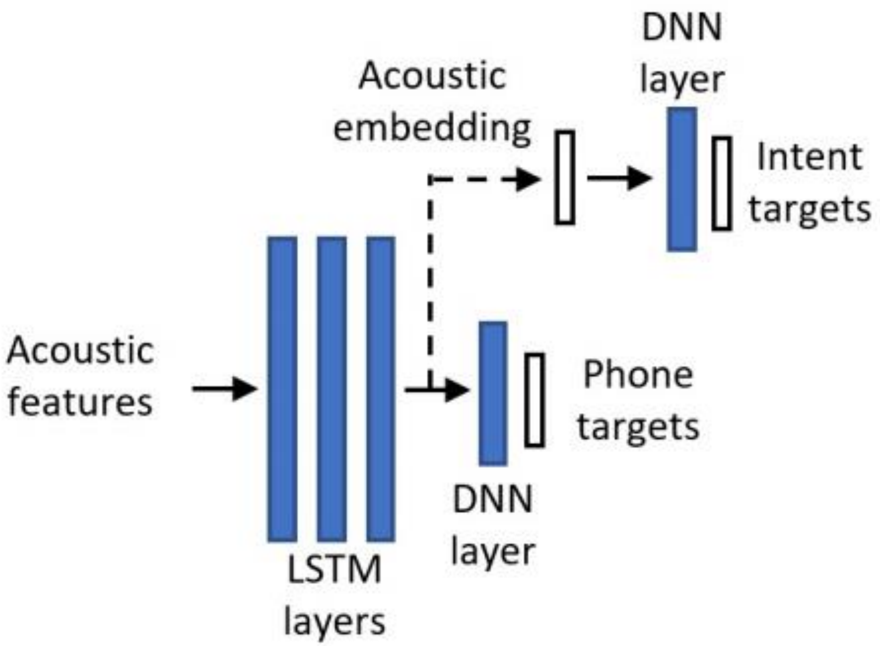


Fig. 1. A S2I system with pre-trained ASR

Leveraging pre-trained networks

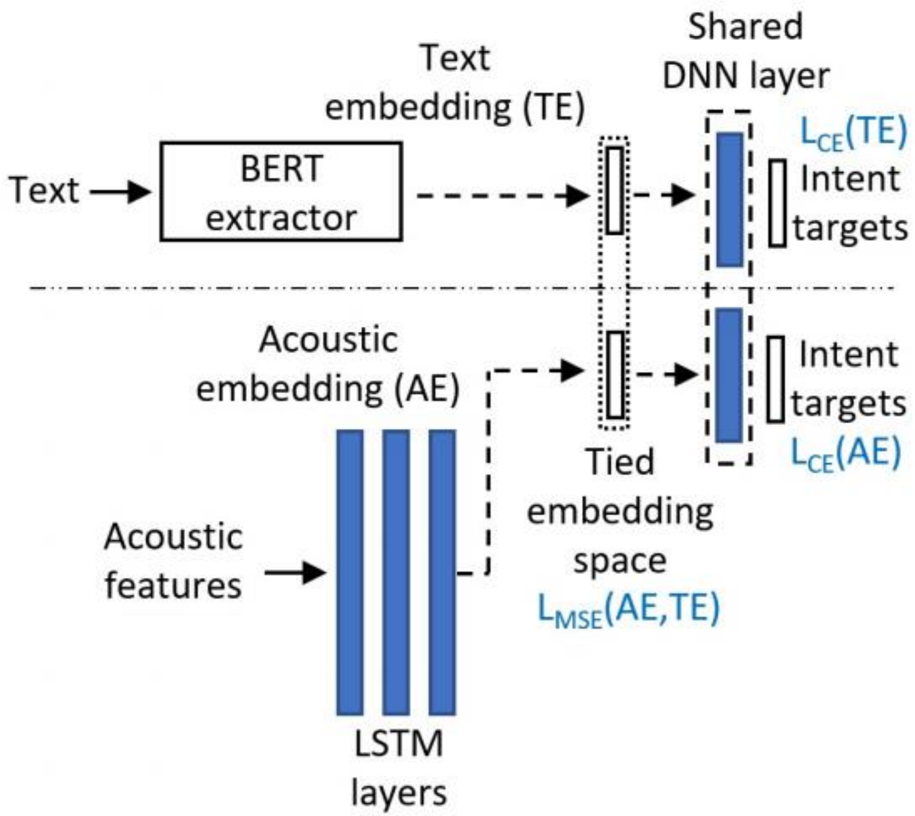


Fig. 2. Joint-training of the S2I system with text embeddings

Leveraging pre-trained networks

IBM

Method	IntAcc
E2E S2I system trained on 2hTrainset	82.2%
Joint training tying speech/text embeddings	84.7%
E2E S2I system trained on 20hTrainset	89.8%

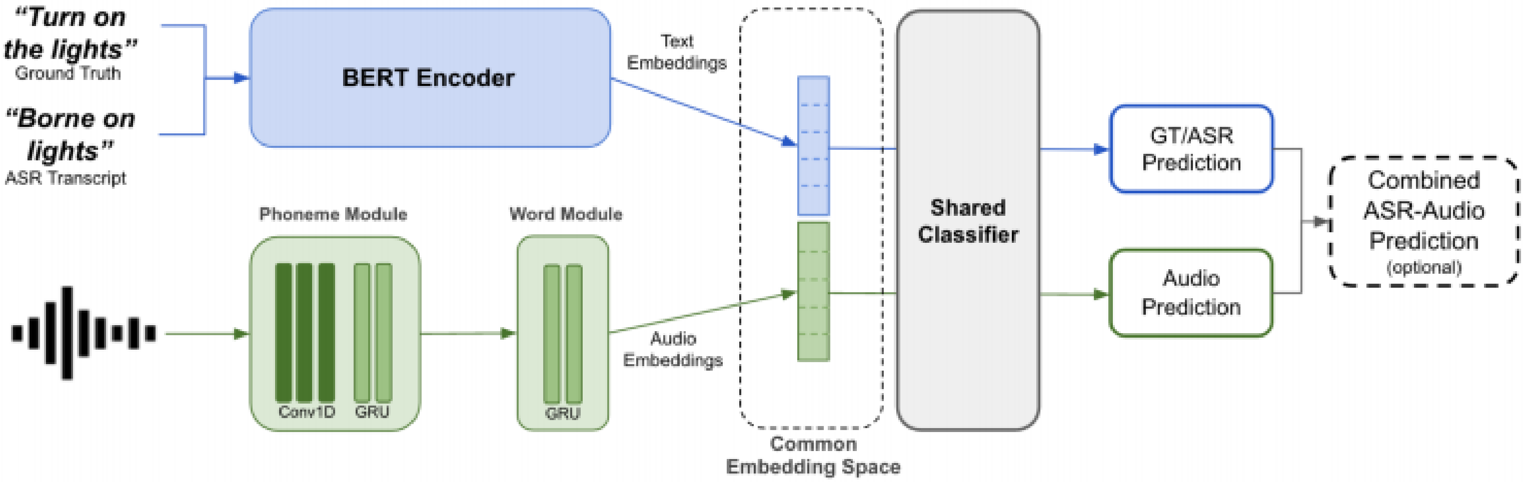
End-to-End models using extra text-to-intent data to recover accuracy lost by switching from *20hTrainset* to *2hTrainset*.

LEVERAGING UNPAIRED TEXT DATA FOR TRAINING END-TO-END SPEECH-TO-INTENT SYSTEMS

Yinghui Huang, Hong-Kwang Kuo, Samuel Thomas, Zvi Kons[†]
Kartik Audhkhasi, Brian Kingsbury, Ron Hoory[†], *Michael Picheny*[*]

IBM Research AI, Yorktown Heights, USA
[†]IBM Research AI, Haifa, Israel

Leveraging pre-trained networks



Speak or Chat with Me:
End-to-End Spoken Language Understanding System with Flexible Inputs

Sujeong Cha[1*], *Wangrui Hou*[1*], *Hyun Jung*[1*], *My Phung*[1*], *Michael Picheny*[1],
Hong-Kwang Kuo[2], *Samuel Thomas*[2], *Edmilson Morais*[3]

[1]New York University, USA
[2]IBM Research AI, USA [3]IBM Research AI, Brazil

SLU as an ASR customization process

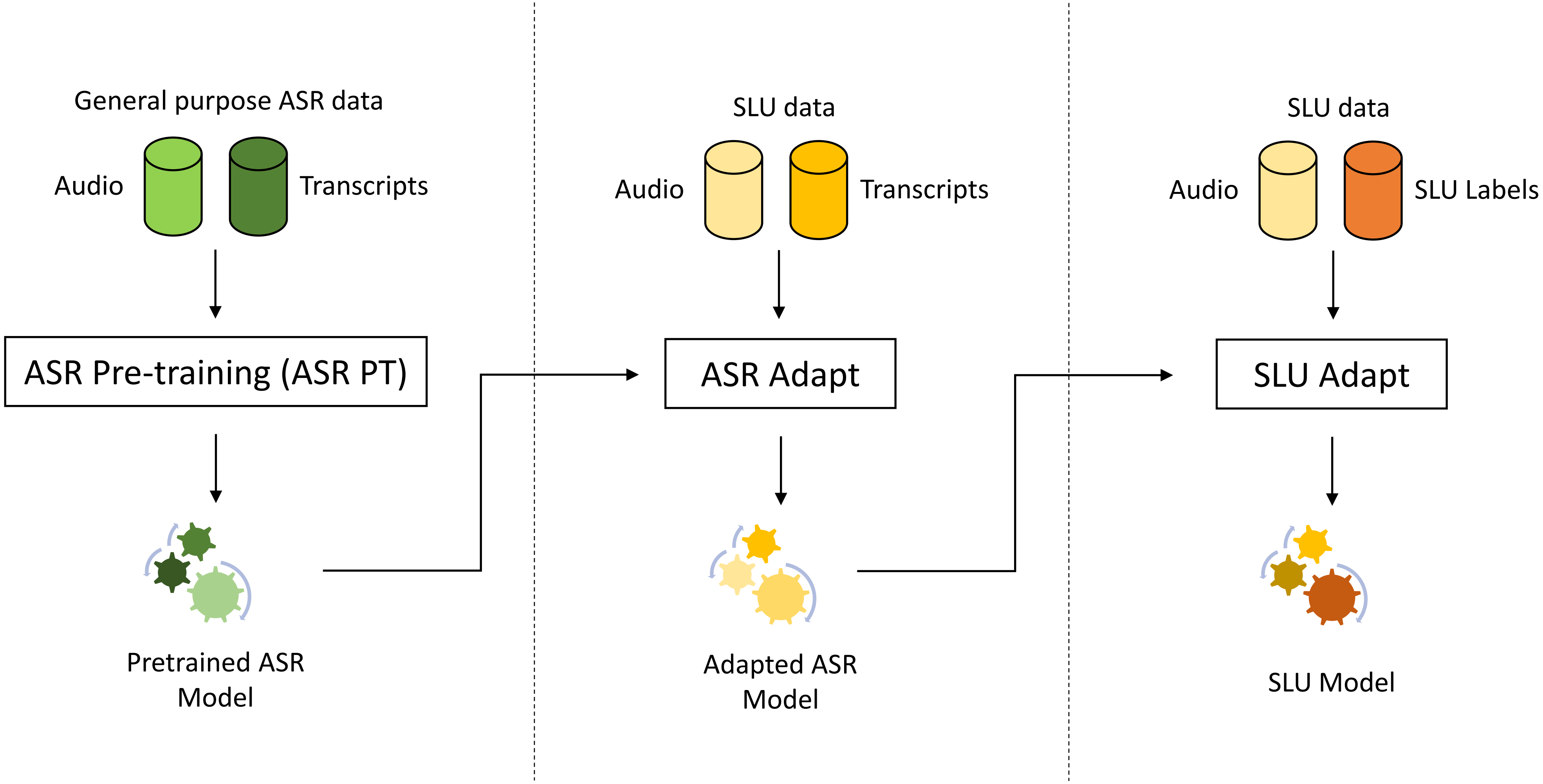
- Directly process speech to produce spoken language understanding (SLU) entity or intent label targets.
  - <speech> I want a flight to Delhi from Chennai that makes a stop in Mumbai
  - **<SLU> Transcript + Intent label**: I want a flight to Delhi from Chennai that makes a stop in Mumbai INT-FLIGHT
  - **<SLU> Transcript + Entity labels**: I want a flight to DELHI B-toloc.cityname from CHENNAI B-fromloc.cityname that makes a stop in MUMBAI B-stoploc.cityname
  - **<SLU> Entity labels only**: DELHI B-toloc.cityname CHENNAI B-fromloc.cityname MUMBAI B-stoploc.cityname
  - **<SLU> Intent label only**: INT-FLIGHT
- Approach the training of SLU models as a kind of **ASR customization process**
  - Start from a pre-trained automatic speech recognition (ASR) system, followed by an SLU adaptation step
  - SLU scenarios
    - a case where verbatim transcripts are available,
    - a constrained case where the only available annotations are SLU labels and their values,
    - a more restrictive case where transcripts are available but not corresponding audio.

Leveraging pre-trained ASR networks

FULL - Speech data is available with transcripts annotated with various SLU labels

(1) Transcript: i want a flight to Delhi from Chennai that makes a stop in Mumbai
(2) Transcript + Entity labels: I want a flight to DELHI B-toloc.cityname from CHENNAI B-fromloc.cityname that makes a stop in MUMBAI B-stoploc.cityname
(3) Transcript + Intent label: i want a flight to Delhi from Chennai that makes a stop in Mumbai INT-FLIGHT

Leveraging pre-trained ASR networks



Leveraging pre-trained ASR networks

How important is have a pre-trained ASR model? How accurate should the pretrained model be?

Table 1: ASR WER performance before and after SLU adaptation. **[P]** denotes experiments focused on pre-training.

	PT. Data (Hrs.)	**ATIS (WER%)**
[1P]	0	14.8
[2P]	64	$38.3 \rightarrow 2.2$
[3P]	160	$18.6 \rightarrow 1.8$
[4P]	300	$13.1 \rightarrow 1.6$

Table 2: SLU performance with various pre-trained models.

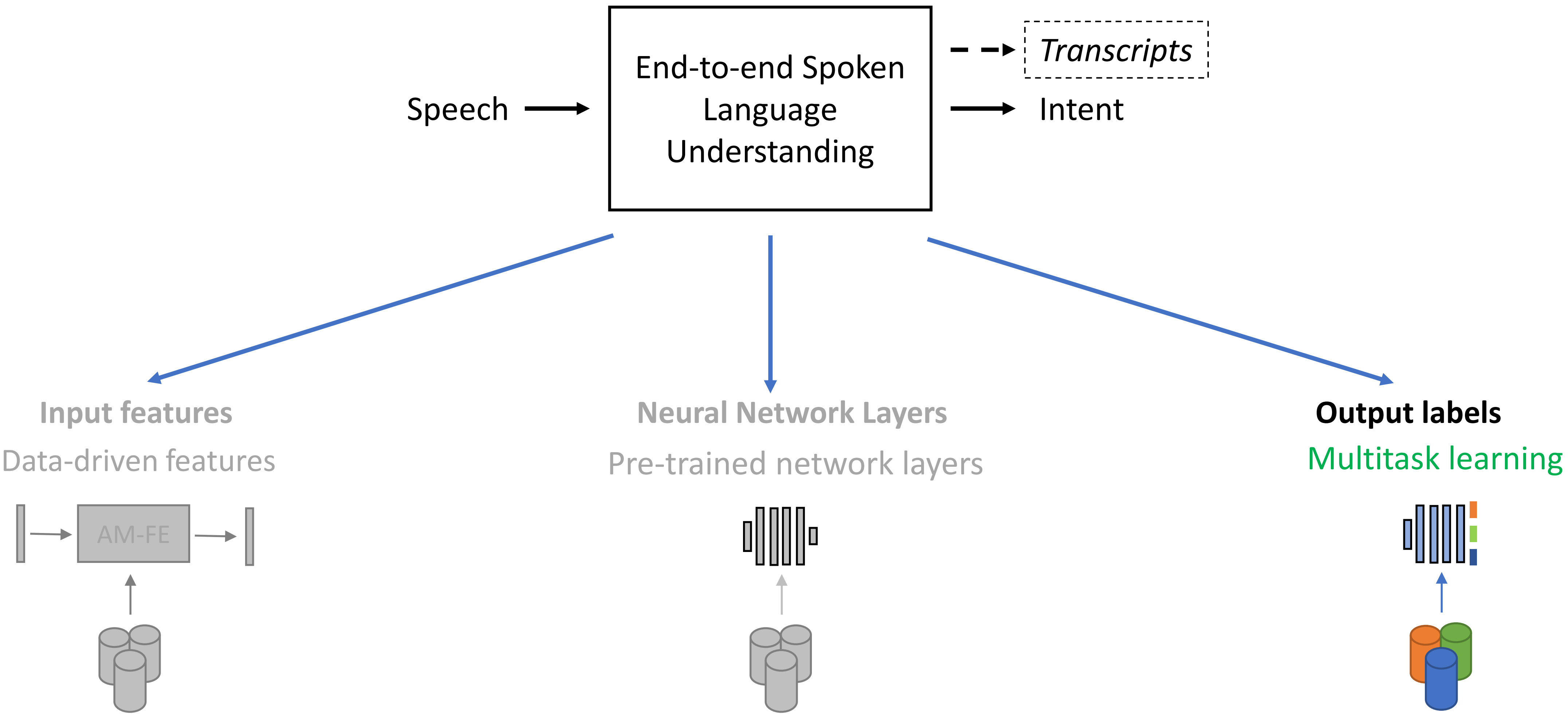
	PT.	**ATIS**		**CC**
	Data (Hrs.)	**Ent. (F1)**	**Int. (Acc%)**	**Int. (Acc%)**
[5P]	0	79.7	83.5	65.8
[6P]	64	92.1	95.4	86.9
[7P]	160	93.2	94.7	87.4
[8P]	300	93.2	94.9	87.4

RNN TRANSDUCER MODELS FOR SPOKEN LANGUAGE UNDERSTANDING

Samuel Thomas, Hong-Kwang J. Kuo, George Saon, Zoltán Tüske,
Brian Kingsbury, Gakuto Kurata, Zvi Kons, Ron Hoory

IBM Research AI

Can we use what we learnt for SLU?



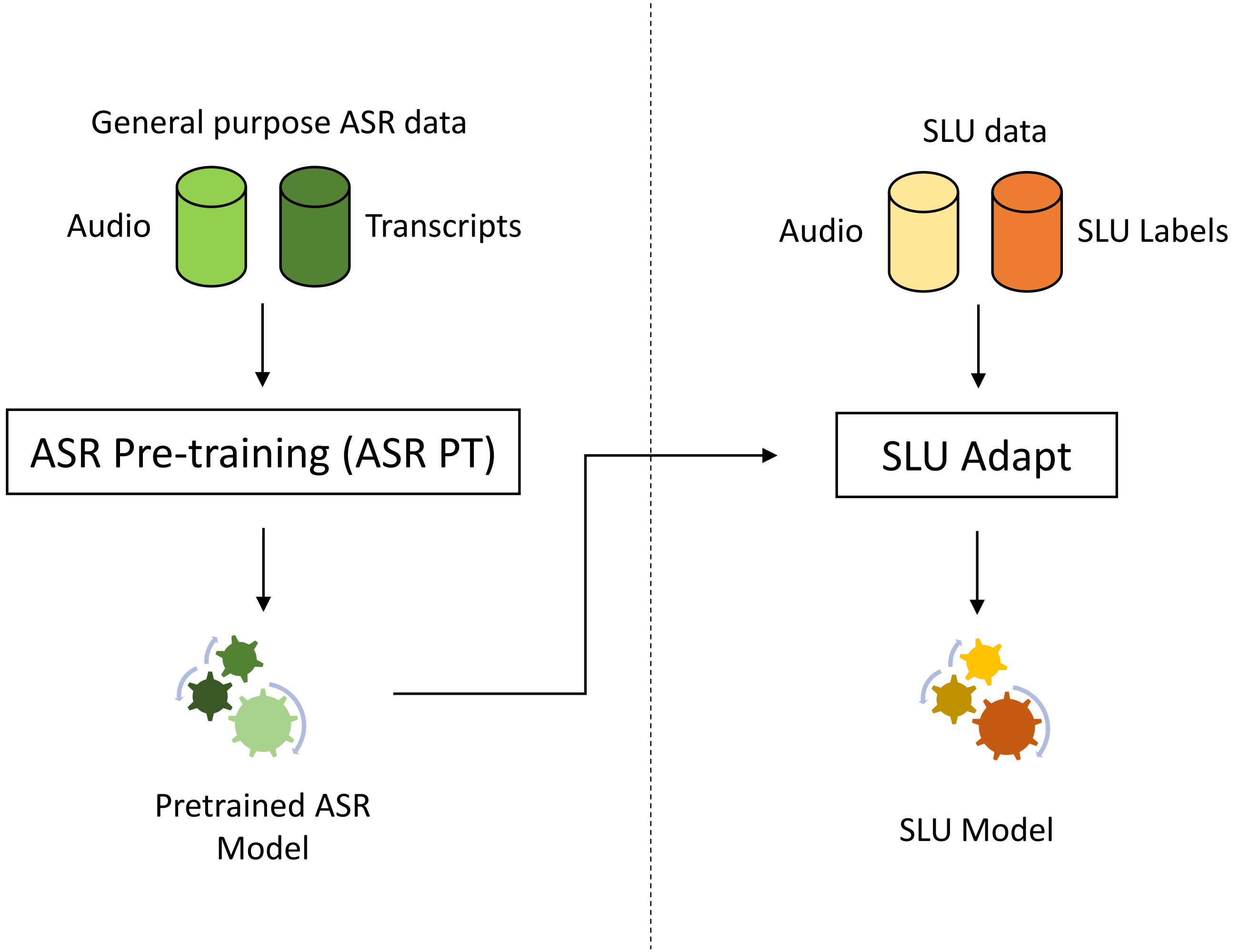
Limited labels

AUDIO - Audio recordings are available, but the annotations are just SLU entity label/value pairs and intents

I want a flight to Delhi from Chennai that makes a stop in Mumbai

(1) Entities in spoken order: DELHI B-toloc.cityname CHENNAI B-fromloc.cityname MUMBAI B-stoploc.cityname
(2) Entities in alphabetic order: CHENNAI B-fromloc.cityname MUMBAI B-stoploc.cityname DELHI B-toloc.cityname
(3) Intent label only: INT-FLIGHT

Limited labels



Limited labels

Training Data	Adapt	CTC	Attention
[1A] Full transcripts	Y	91.7	92.9
[2A] Full transcripts	N	91.7	93.0
[3A] Entities, spoken order	Y	92.7	92.8
[4A] Entities, spoken order	N	91.5	92.6
[5A] Entities, alphabetic order	Y	73.5	90.9
[6A] Entities, alphabetic order	N	61.9	90.6

ATIS *bag-of-entities slot filling F1 score for speech input using CTC and Attention based models*

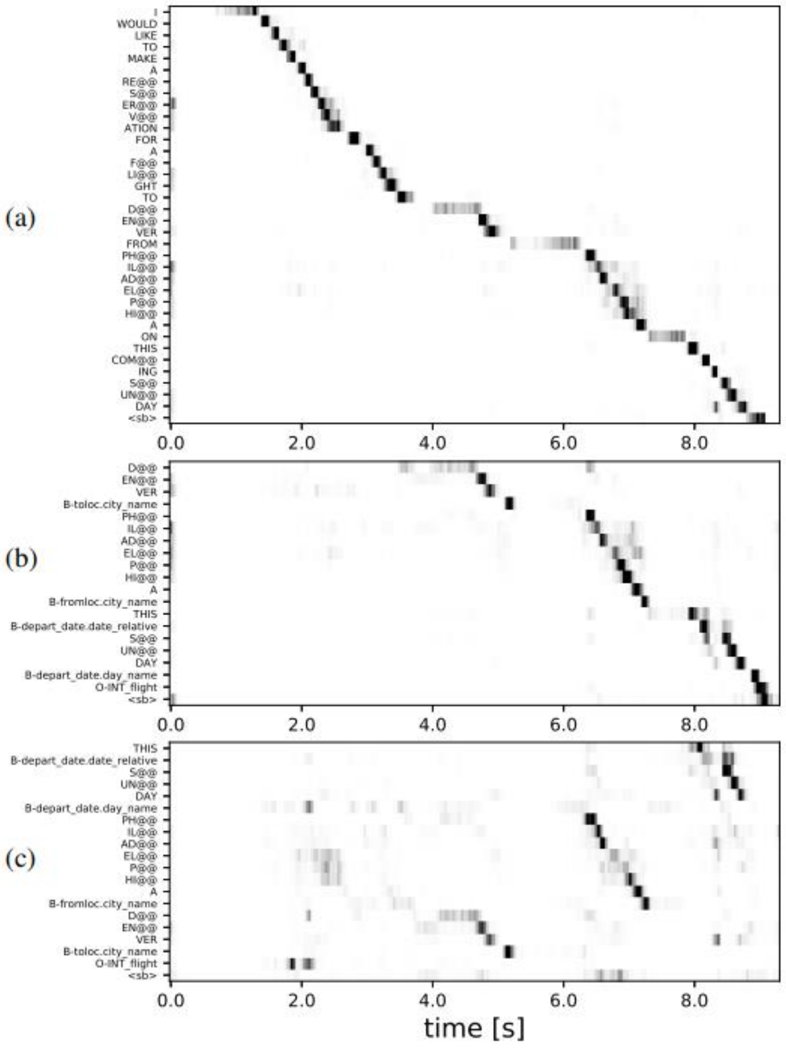


Figure 1: *Attention plots for the utterance "I would like to make a reservation for a flight to Denver from Philadelphia on this coming Sunday": (a) ASR; (b) SLU in spoken order; (c) SLU in alphabetic order.*

Limited labels

Training Data	Adapt	CTC	Attention
[1B] Full transcripts	Y	85.5	92.0
[2B] Full transcripts	N	79.6	91.3
[3B] Entities, spoken order	Y	88.6	91.2
[4B] Entities, spoken order	N	86.5	89.6
[5B] Entities, alphabetic order	Y	73.8	88.8
[6B] Entities, alphabetic order	N	68.5	87.7

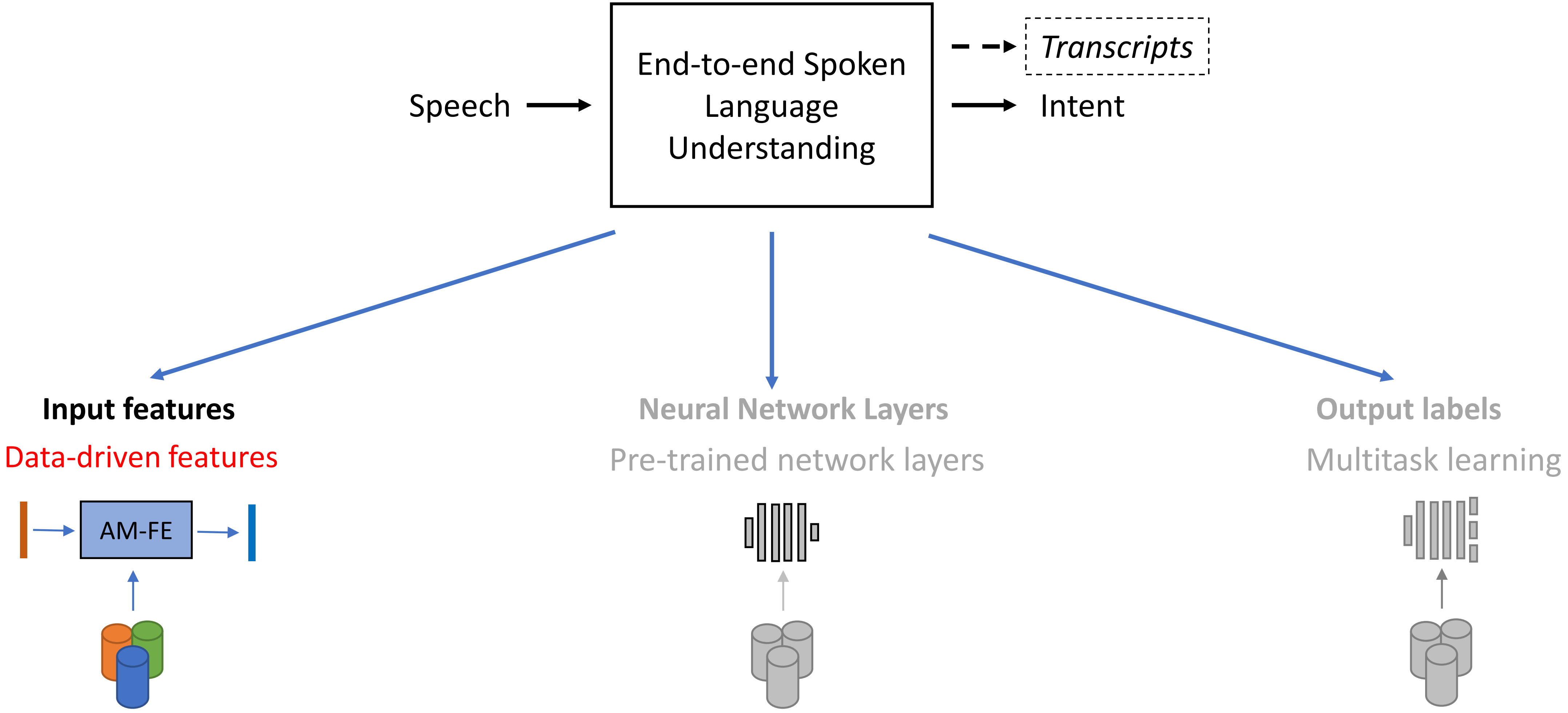
ATIS *bag-of-entities slot filling F1 score for speech input with additive street noise (5dB SNR)*

End-to-End Spoken Language Understanding Without Full Transcripts

Hong-Kwang J. Kuo, Zoltán Tüske, Samuel Thomas, Yinghui Huang, Kartik Audhkhasi*, Brian Kingsbury, Gakuto Kurata, Zvi Kons, Ron Hoory, and Luis Lastras*

IBM Research AI

Can we use what we learnt for SLU?



Features and data augmentation

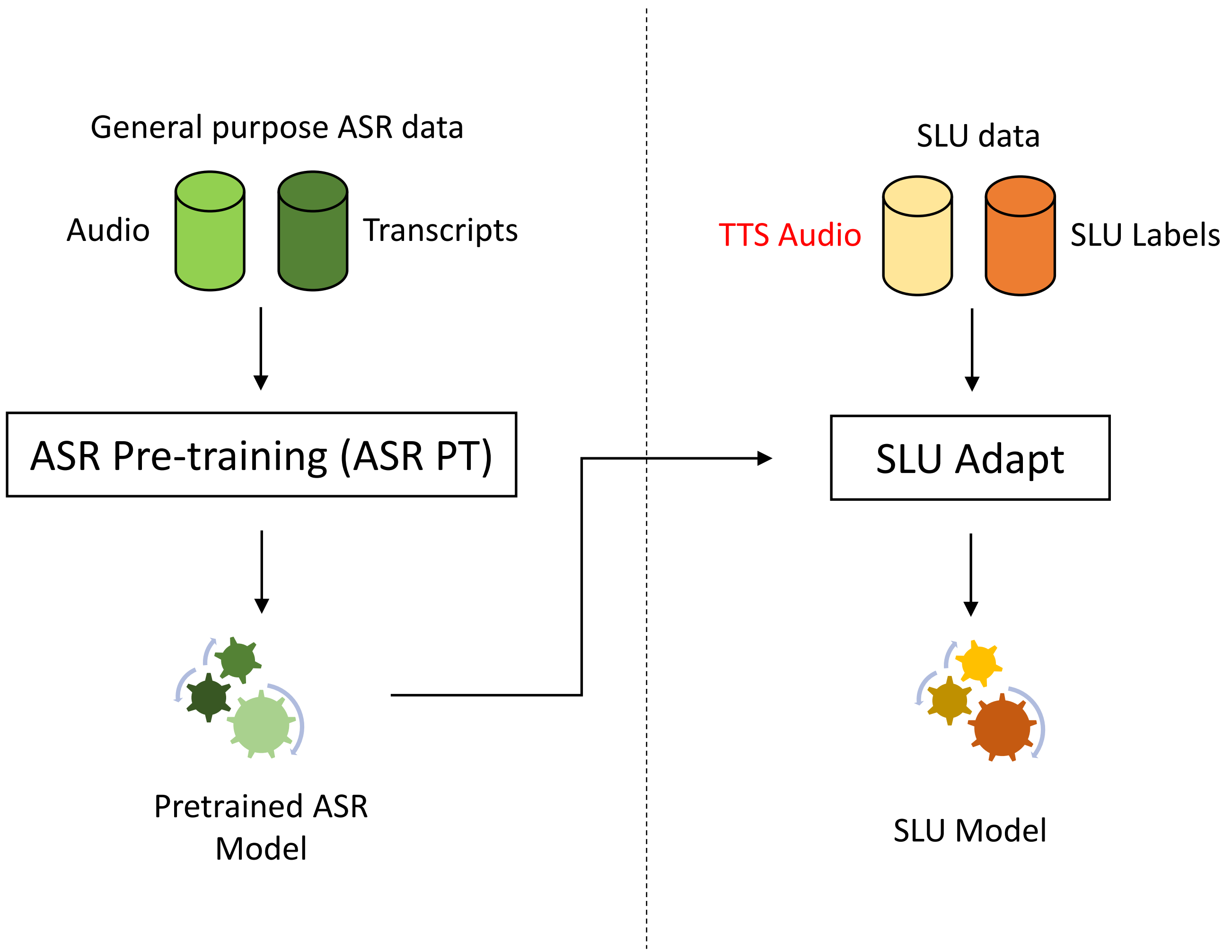
IBM

TEXT - Transcripts with SLU annotations are available, but the corresponding human speech recordings are not, due to privacy restrictions or bootstrapping from text chat data.

(1) Transcript + Entity labels: I want a flight to DELHI B-toloc.cityname from CHENNAI B-fromloc.cityname that makes a stop in MUMBAI B-stoploc.cityname

(2) Transcript + Intent label: I want a flight to Delhi from Chennai that makes a stop in Mumbai INT-FLIGHT

Features and data augmentation

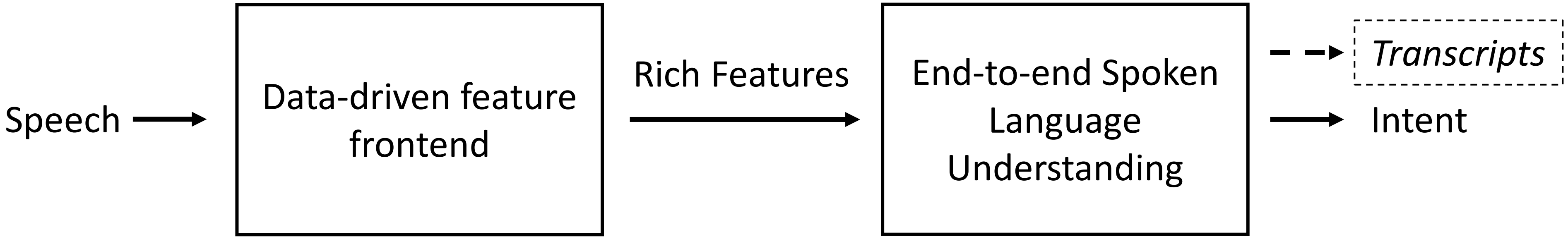


Features and data augmentation

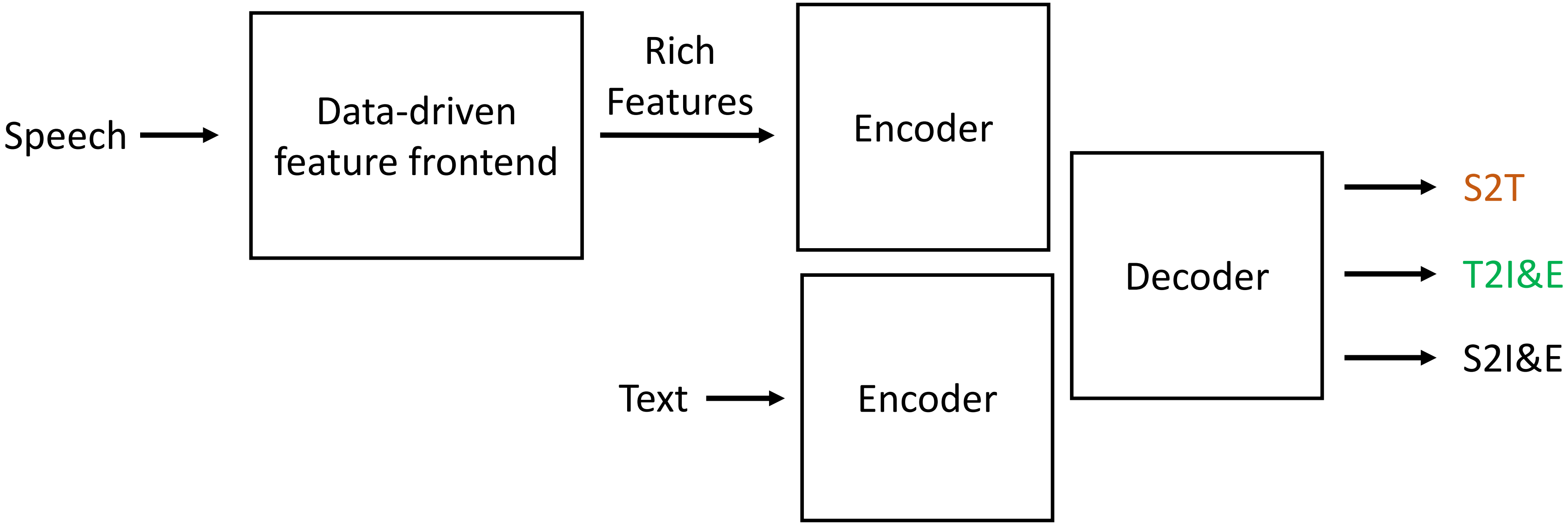
Method	IntAcc
E2E S2I system trained on 2hTrainset	82.2%
Joint training tying speech/text embeddings	84.7%
Adding synthetic multi-speaker TTS speech	87.8%
Joint training + adding synthetic speech	88.3%
E2E S2I system trained on 20hTrainset	89.8%

End-to-End models using extra text-to-intent data to recover accuracy lost by switching from *20hTrainset* to *2hTrainset*.

Features and data augmentation



Features and data augmentation



END-TO-END SPOKEN LANGUAGE UNDERSTANDING USING TRANSFORMER NETWORKS AND SELF-SUPERVISED PRE-TRAINED FEATURES

Edmilson Morais, Hong-Kwang J. Kuo, Samuel Thomas, Zoltán Tüske and Brian Kingsbury

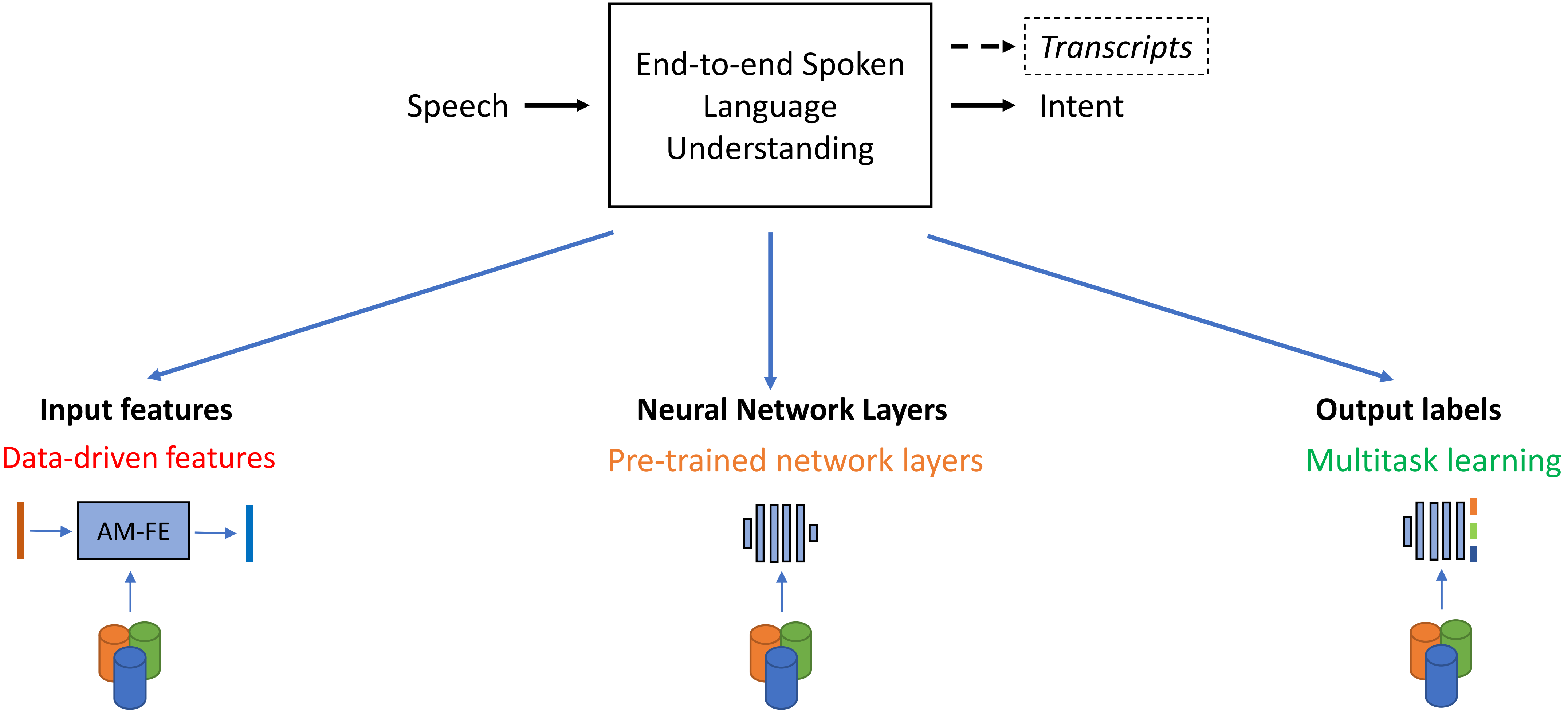
IBM Research

Features and data augmentation

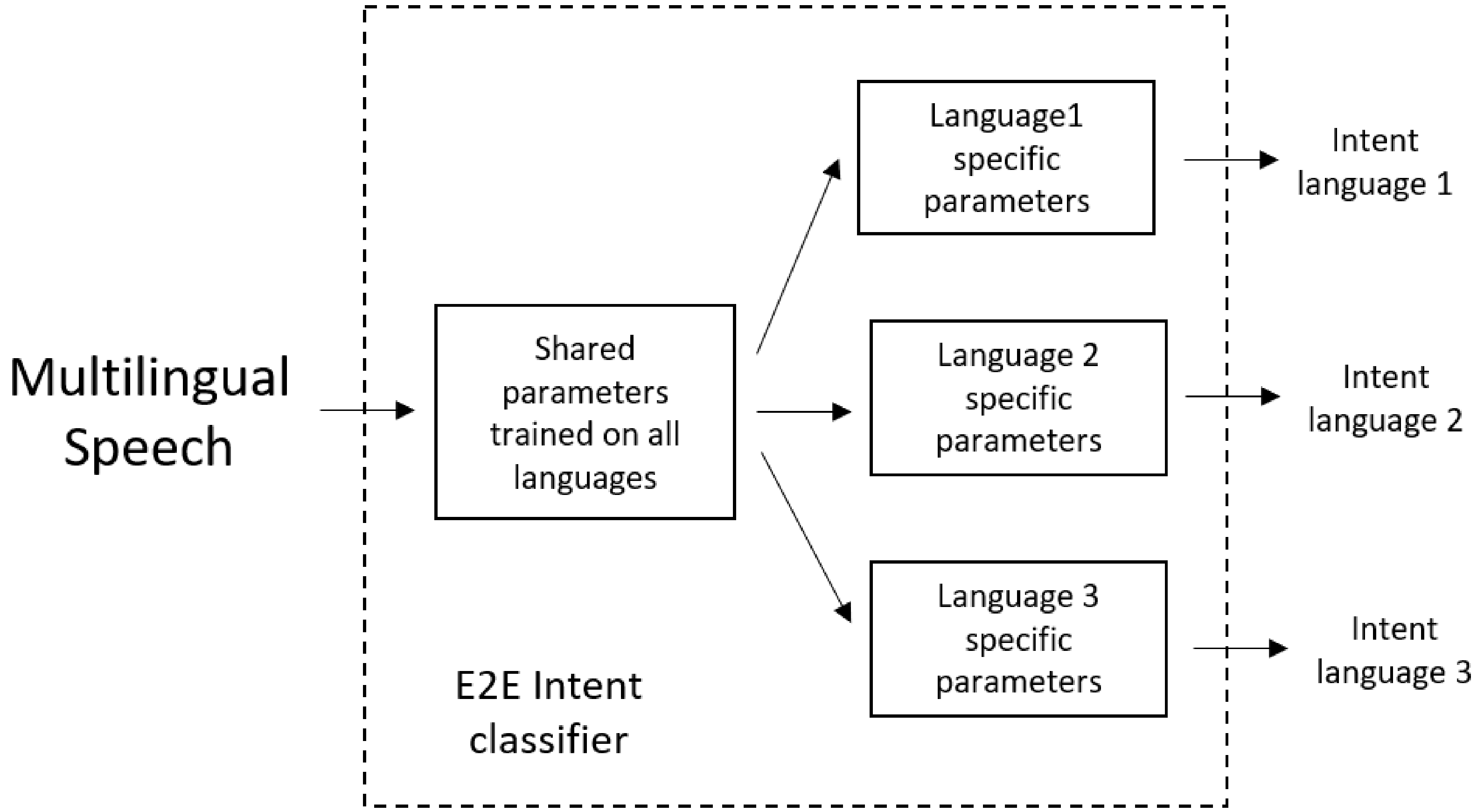
Number	Pre-initialization		Auxiliary tasks		Entities (F1 score %)		Intent (IER %)	
	Encoder	Decoder	S2T	T2I&E	Filterbank	Wav2vec	Filterbank	Wav2vec
1	-	-	-	-	34.4	76.6	14.0	6.8
2	-	-	-	yes	53.0	83.6	13.0	4.8
3	-	-	yes	-	75.1	89.1	8.5	3.9
4	-	-	yes	yes	**87.0**	**89.3**	**5.8**	**3.3**
5	ATIS	ATIS	-	-	88.1	90.0	3.5	3.4
6	ATIS	ATIS	yes	yes	88.6	91.1	3.8	3.5
7	ATIS	ATIS	yes	-	90.1	**91.4**	3.5	3.3
8	ATIS	-	yes	yes	**91.2**	91.2	**3.4**	**3.3**
9	LibSp100h	-	yes	yes	90.7	89.2	3.3	3.9
10	LibSp100h	ATIS	yes	yes	91.1	**90.0**	3.3	**3.0**

Benefits of pre-training, auxiliary tasks and data driven features
for Entity and Intent Recognition on ATIS

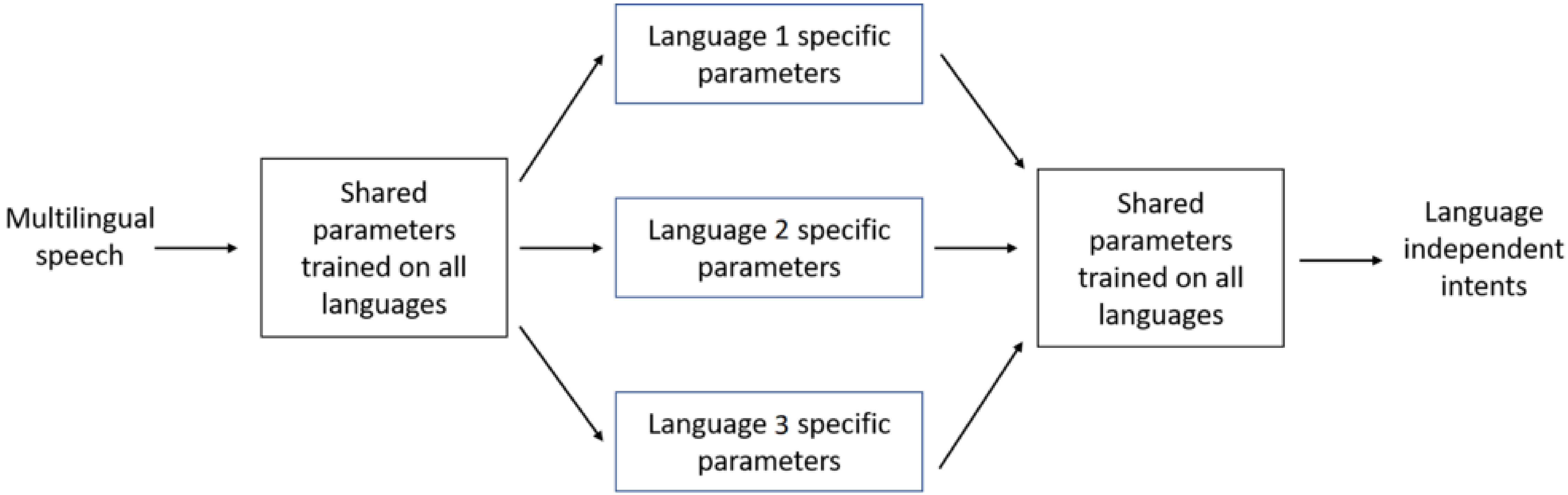
Can we use what we learnt for SLU?



What next? Multilingual SLU



What next? Multilingual SLU



Conclusion

- Discussed the E2E SLU task and showed how various E2E SLU models are trained in a very similar fashion to low-resource multilingual models with
  - Pre-trained models
  - Data driven features
  - Multi-task learning
  - Limited transcripts
- Propose a focus on multilingual E2E SLU and related tasks as a related task with significant value

Acknowledgements

Hong-Kwang J Kuo, George Saon, Zoltán Tüske, Brian Kingsbury,
Gakuto Kurata, Zvi Kons, Ron Hoory, Edmilson Morais, Sujeong Cha,
Wangrui Hou, Hyun Jung, My Phung, Michael Picheny, Yinghui Huang,
Kartik Audhkhasi, and Luis Lastras


THANK YOU