Select On-device Spoken Language Understanding Topics

Jimmy Kunzmann
(Credits: Ariya Rastrow & Björn Hoffmeister & Daniel Willett)
Alexa Speech, Amazon

amazon | science

Alexa ASR Science

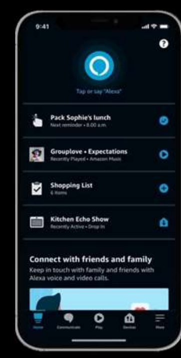
We do In-Cloud, On-Device and In-Car ASR for
- Human-Machine Interactions (e.g., Alexa)
- Human Speech Transcription (e.g., Voice Search)
- Human-Human-Machine Conversations (e.g., Alexa Conversations)

Where we are:

Alexa enabled Products

We build ASR for …

- Headless devices
- Multi-modal devices
- Smart remotes
- Mobile
- Auto
- Wearables
- Robots

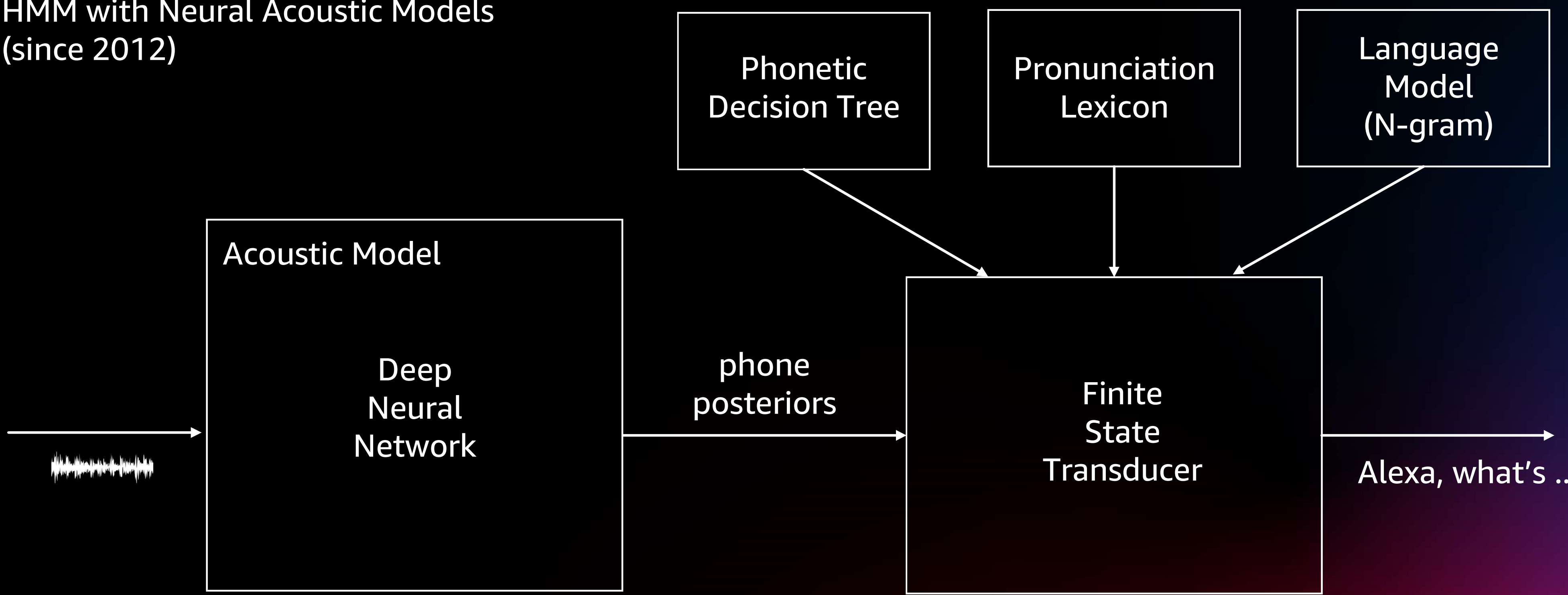

Select On-device Spoken Language Understanding Topics

Agenda

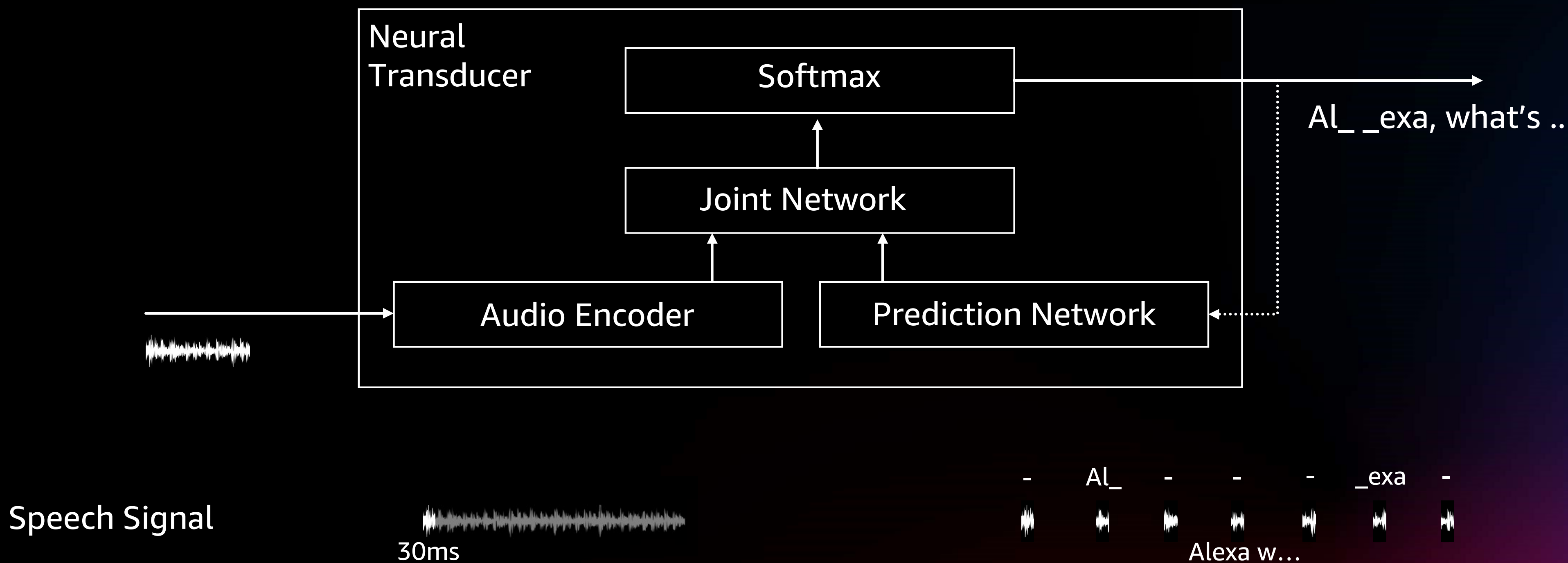
- Birds eye view: Finite State Transducer to Neural Transducer ASR
- Dynamic Adaptation and Personalization
- E2E Speech To Understanding
- Edge Processing – Small Footprint ASR

Finite State Transducer (FST) Based ASR

HMM with Neural Acoustic Models
(since 2012)



Neural Transducer Based ASR



Neural Transducer Based ASR – Pros/Cons

Pros

- End-to-end optimizable
- Representation Learning
- Multi-Task Learning
- (Theoretically) Open Vocabulary
- Accuracy wins

Cons

- Not easy to train
- Expensive to train
  (4-5 weeks on 96 GPUs)
- Rare words are challenging
- Personalization is challenging
- Hotfixing is challenging

H. Tulsiani et al., "Improved training strategies for end-to-end speech recognition in digital voice assistants", Interspeech 2020

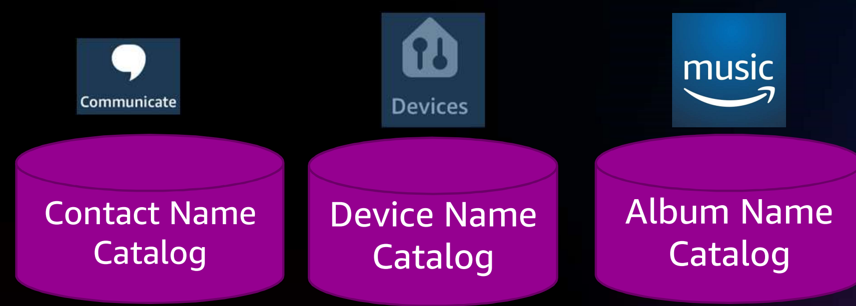
E. Lakomkin et al., "Subword regularization: an analysis of scalability and generalization for end-to-end automatic speech recognition", Interspeech 2022

Dynamic Adaptation and Personalization

- Difficulty Recognizing uncommon/rare words & phrases (All neural models thrive from data)

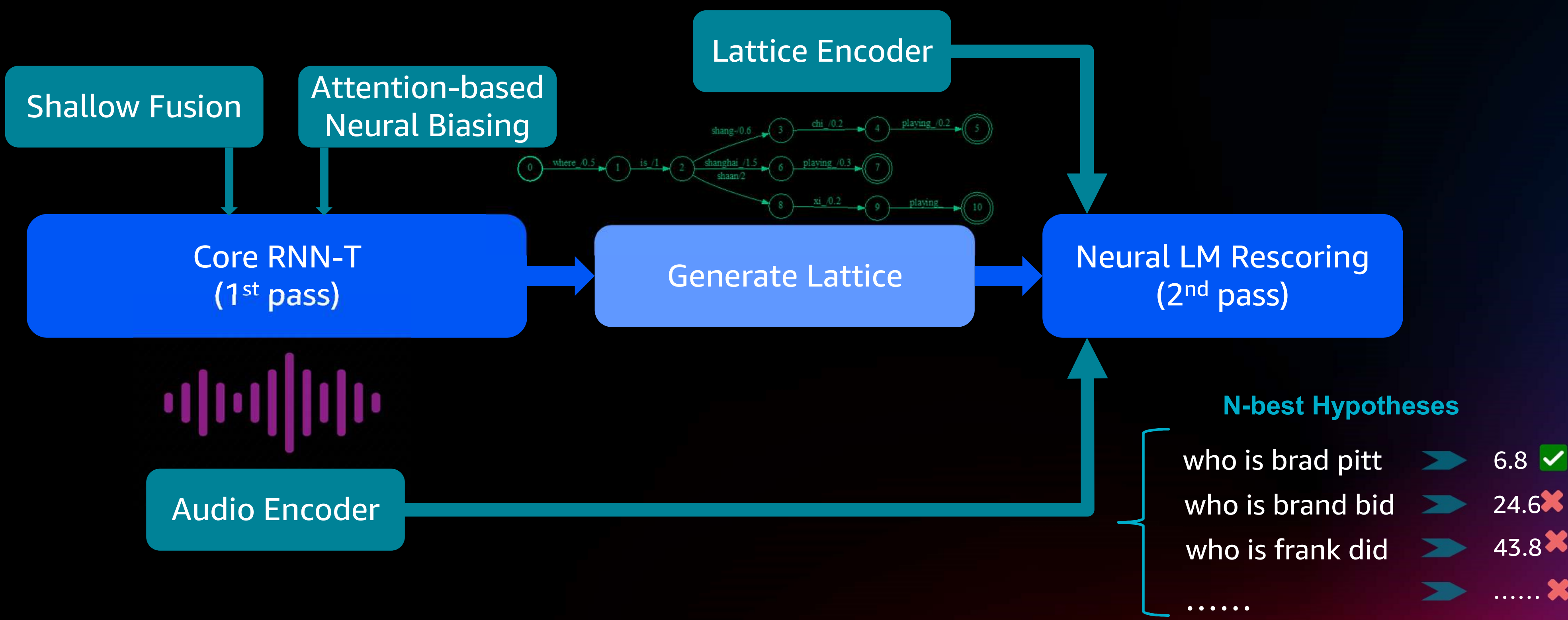
When is movie "X" coming to the theatres?
Call "Y" on his/her cellphone.
Play my "Z" playlist from Spotify.

- Boost personalized entities and catalogs
(ContactNames, PlayList, etc.)



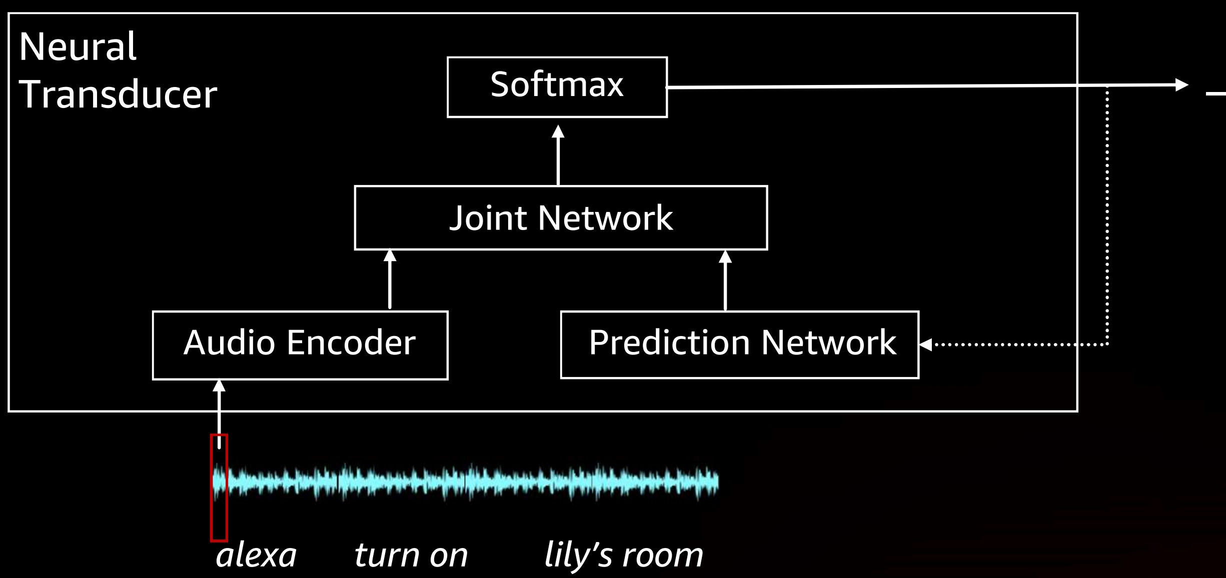
- Domain adaptation
  - Usage shifts overtime
  - Need to support new domains and use cases (cold-start problem) (text-only adaptation)

Dynamic Adaptation and Personalization



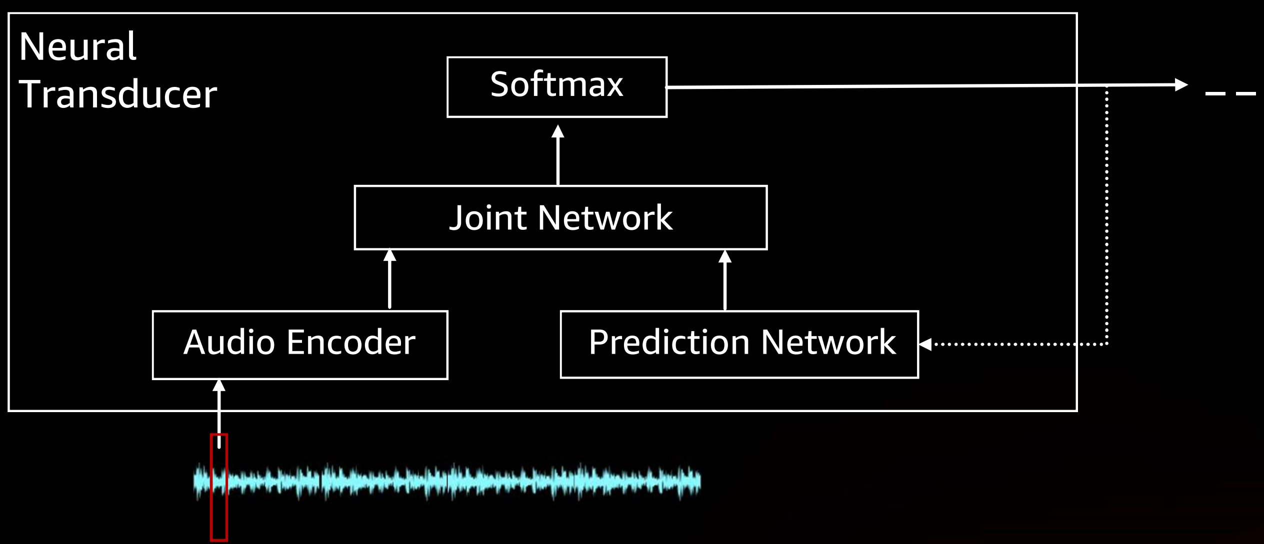
Dynamic Adaptation and Personalization

Attention-based Neural Biasing



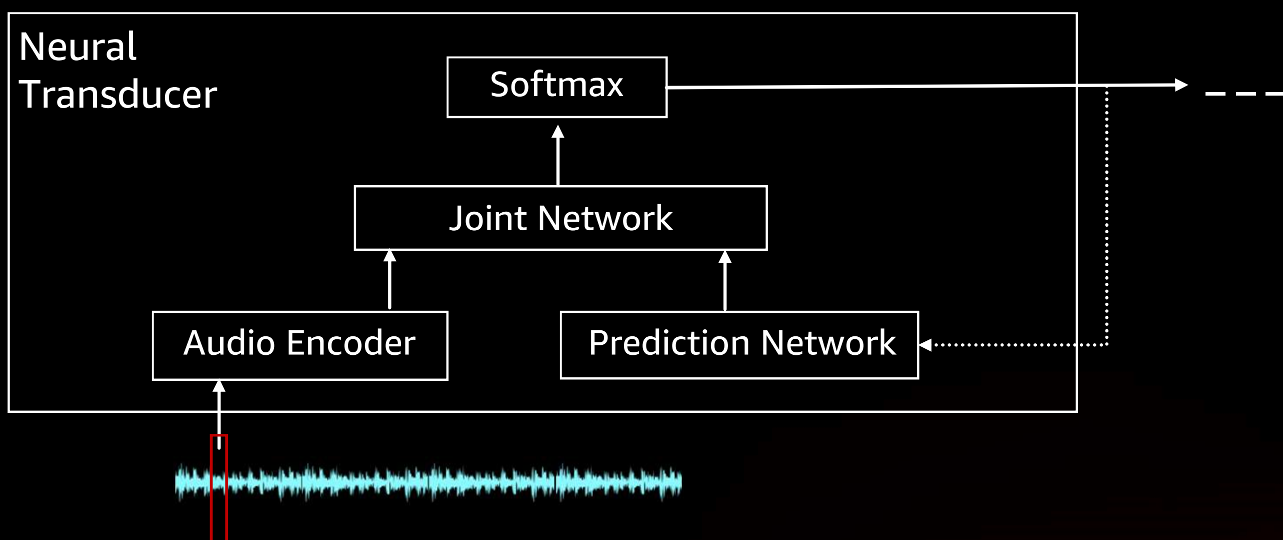
Dynamic Adaptation and Personalization

Attention-based Neural Biasing



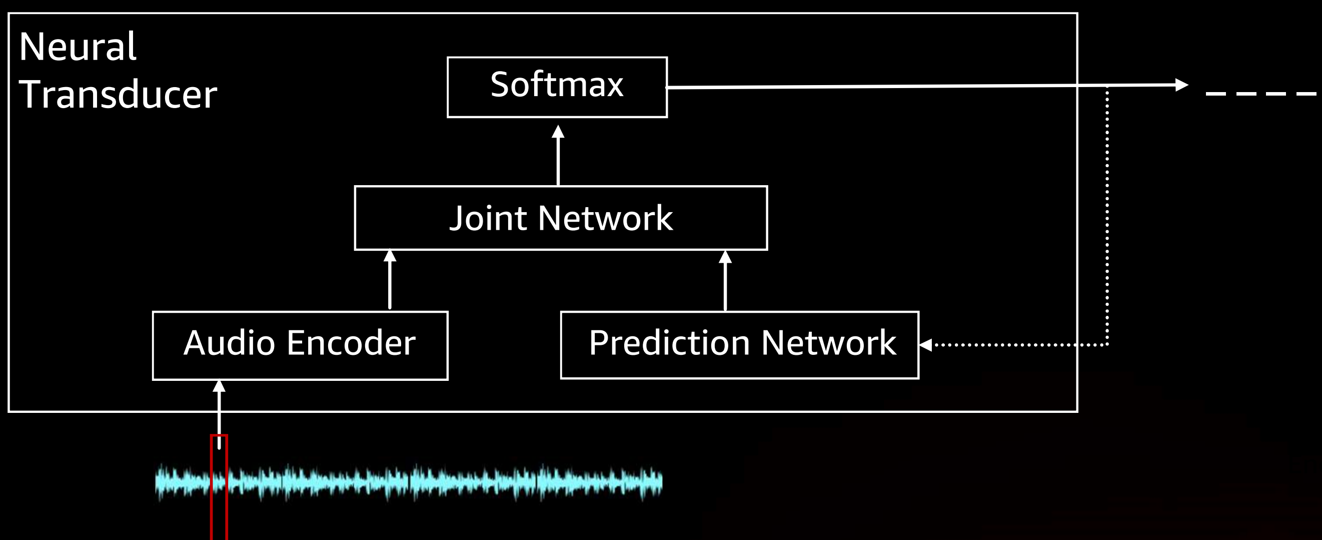
Dynamic Adaptation and Personalization

Attention-based Neural Biasing



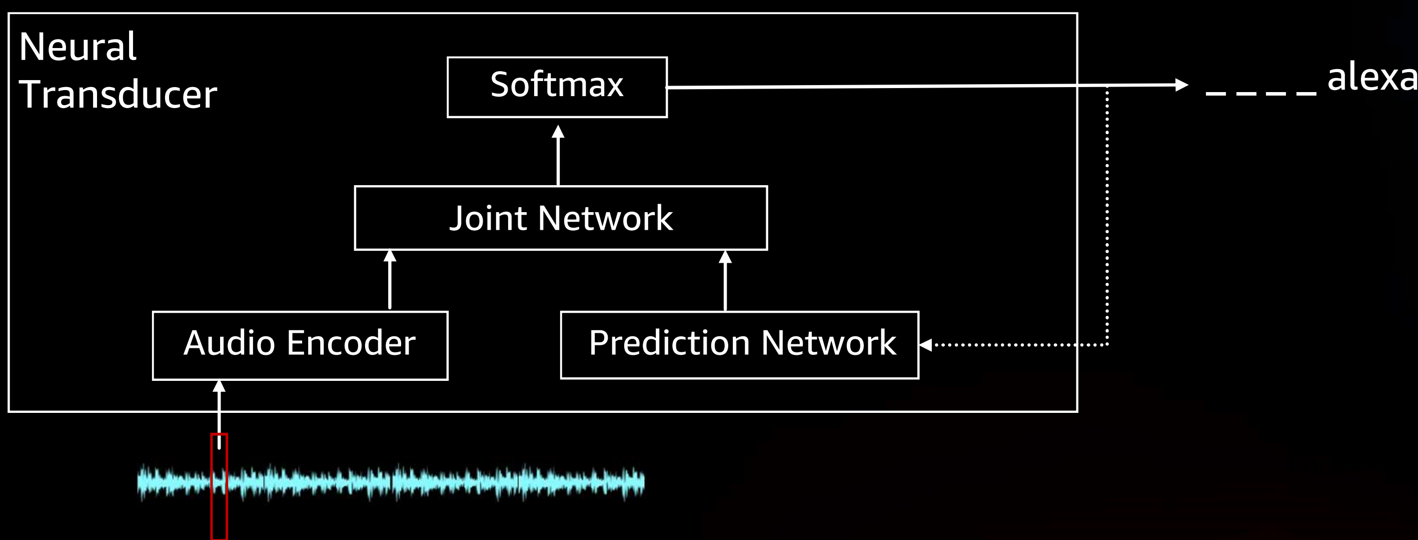
Dynamic Adaptation and Personalization

Attention-based Neural Biasing



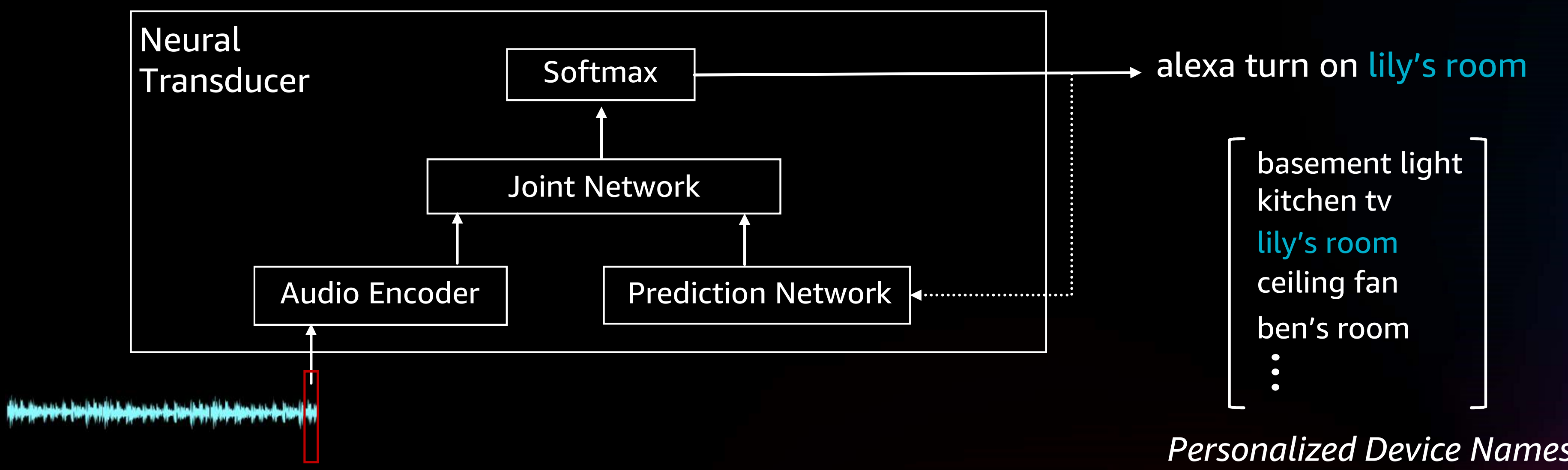
Dynamic Adaptation and Personalization

Attention-based Neural Biasing



Dynamic Adaptation and Personalization

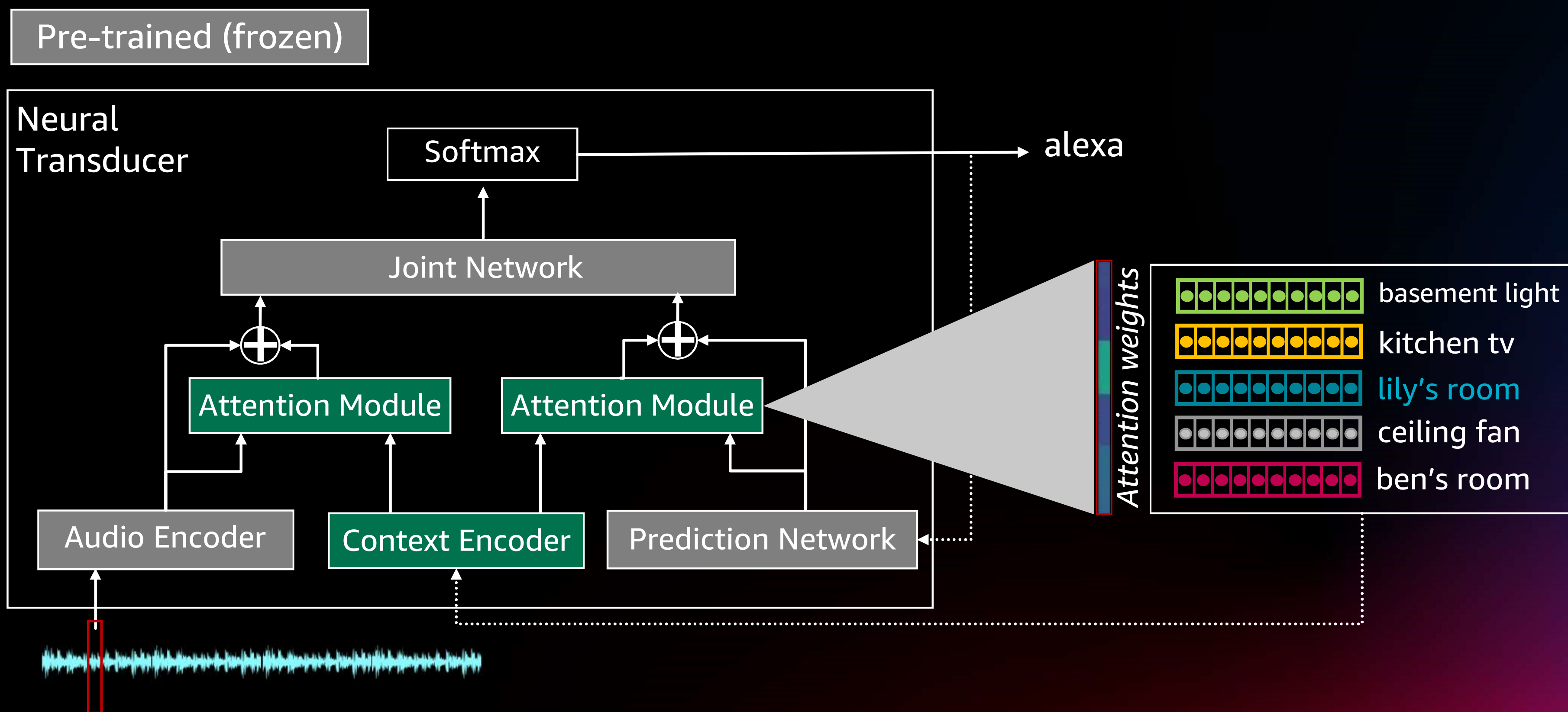
Attention-based Neural Biasing



Personalized Device Names

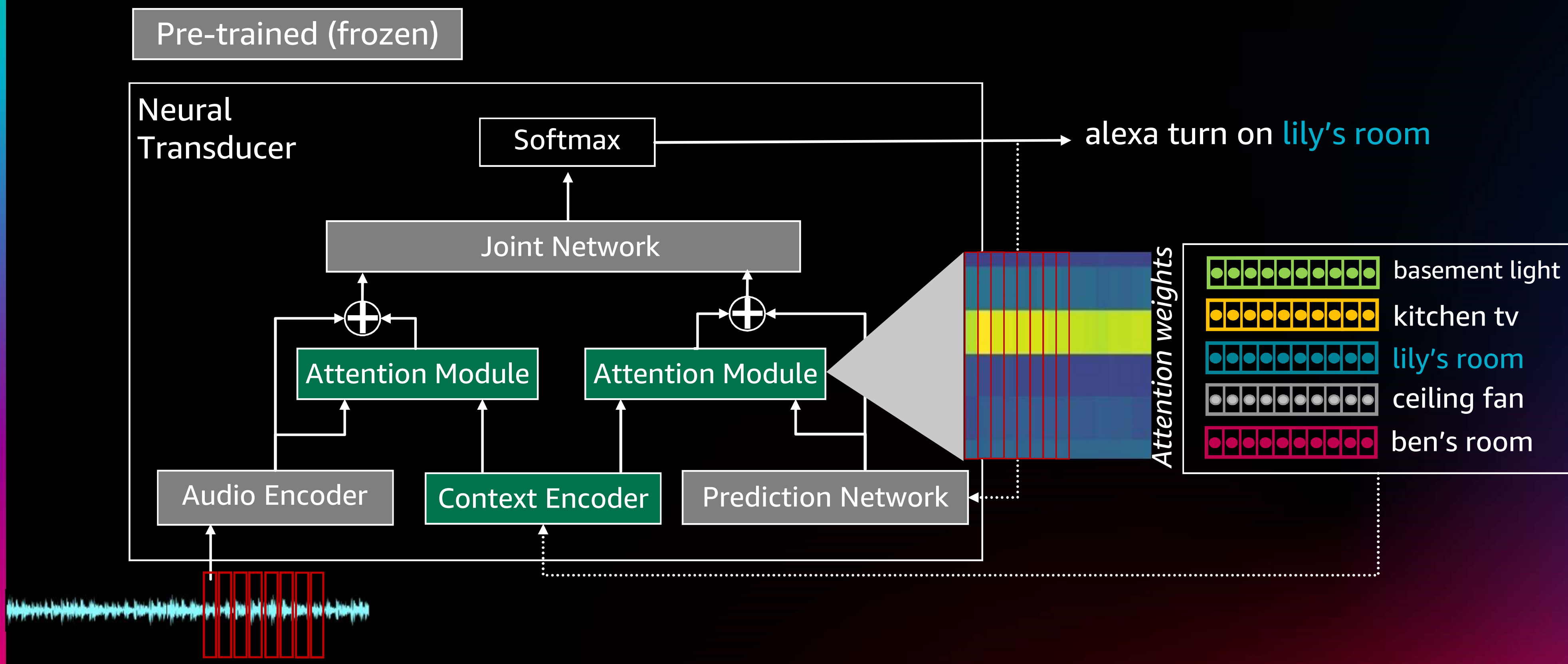
Dynamic Adaptation and Personalization

Attention-based Neural Biasing



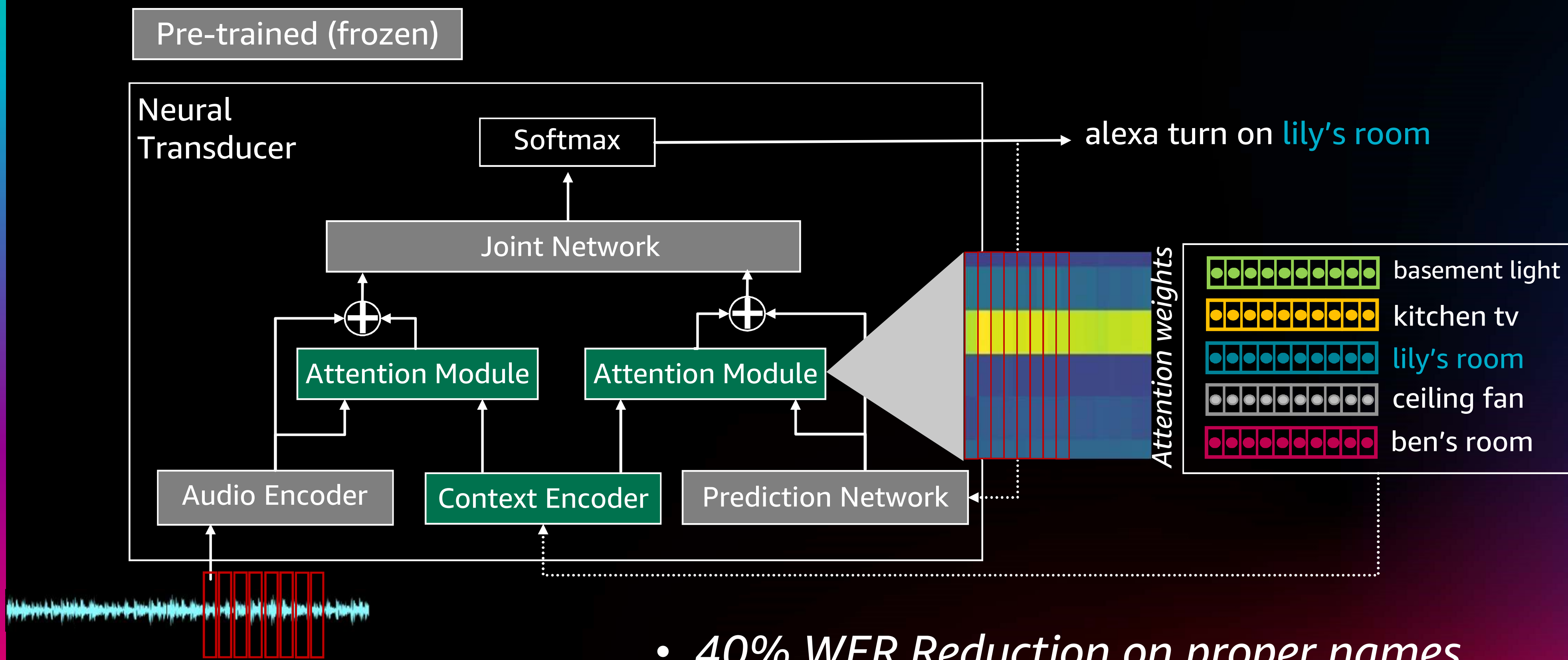
Dynamic Adaptation and Personalization

Attention-based Neural Biasing



Dynamic Adaptation and Personalization

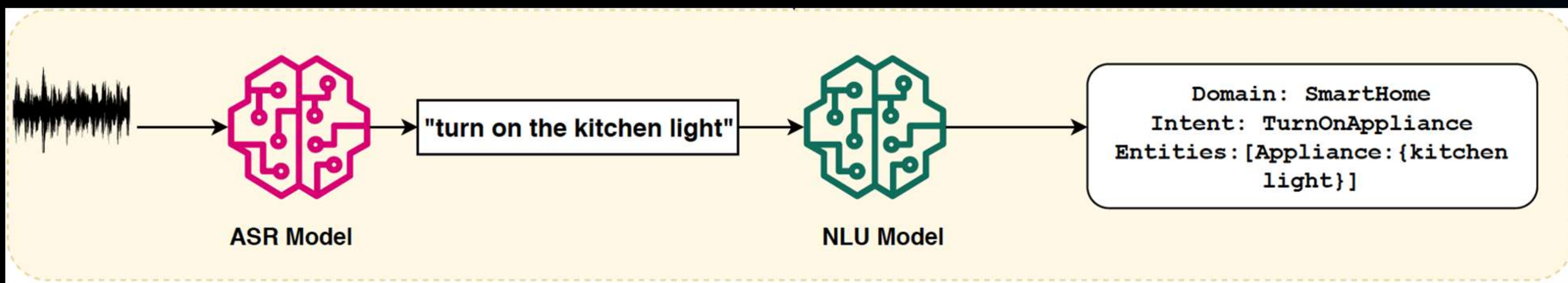
Attention-based Neural Biasing



- *40% WER Reduction on proper names*

E2E Speech To Understanding

Conventional Spoken Language Understanding (SLU) System



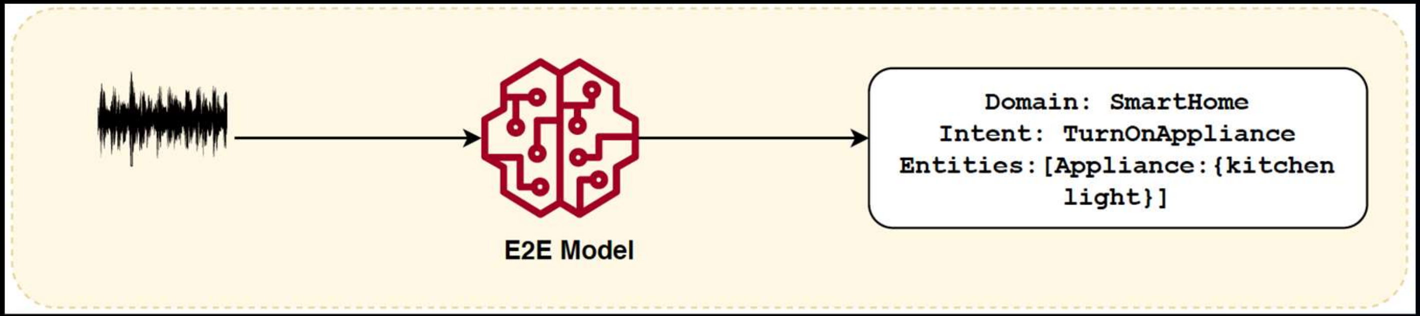
Drawbacks of a Modular SLU System with Independent ASR & NLU Models

- Independent Training
- Training Errors Are Propagated
- Each Error Treated Equally

✗ "turn ~~on~~ the light",
✓ "turn on ~~the~~ light",
✗ "turn on the ~~light~~"

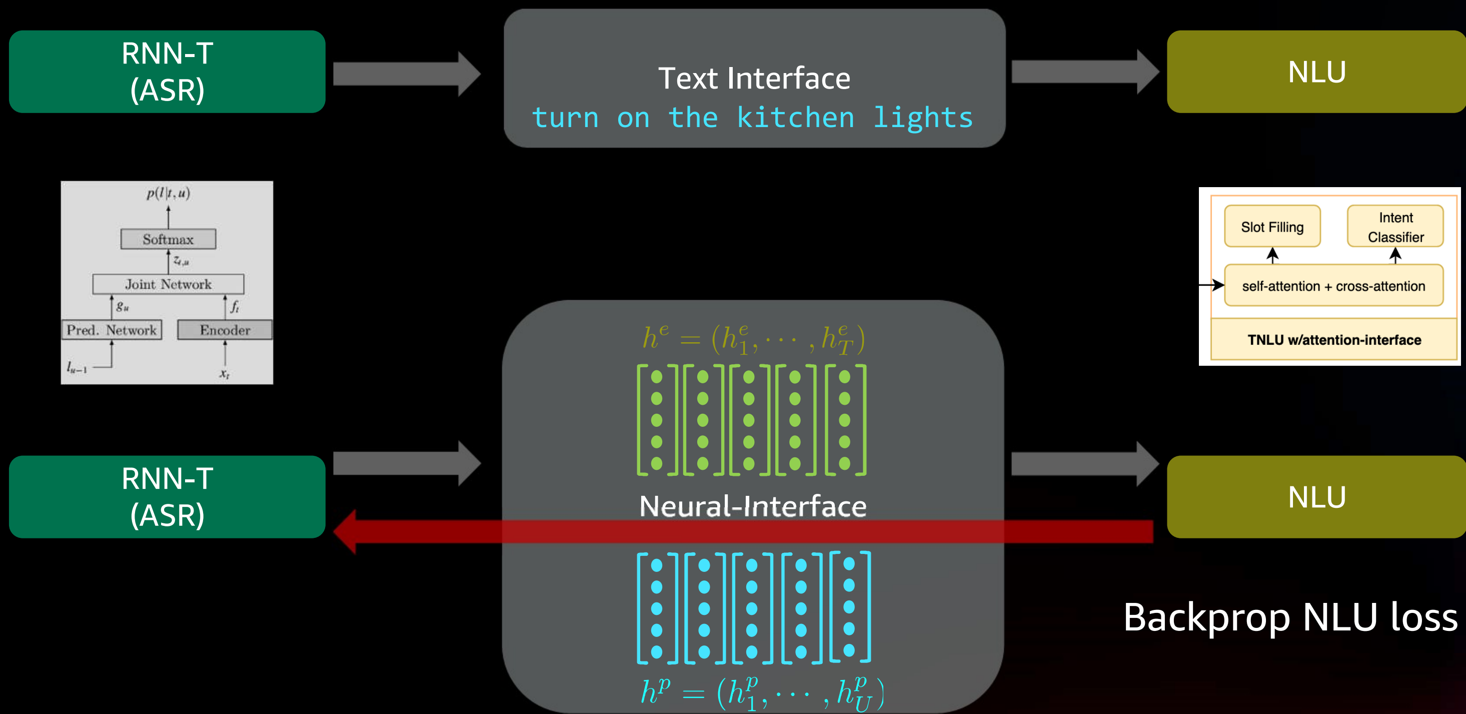
E2E Speech To Understanding

Tighter integration for



- Produce an SLU output directly from the speech signal input
- Either trained with a single optimization objective or jointly optimized end-to-end
- “Error-Robust” as well as “Resource Efficient”

E2E Speech To Understanding

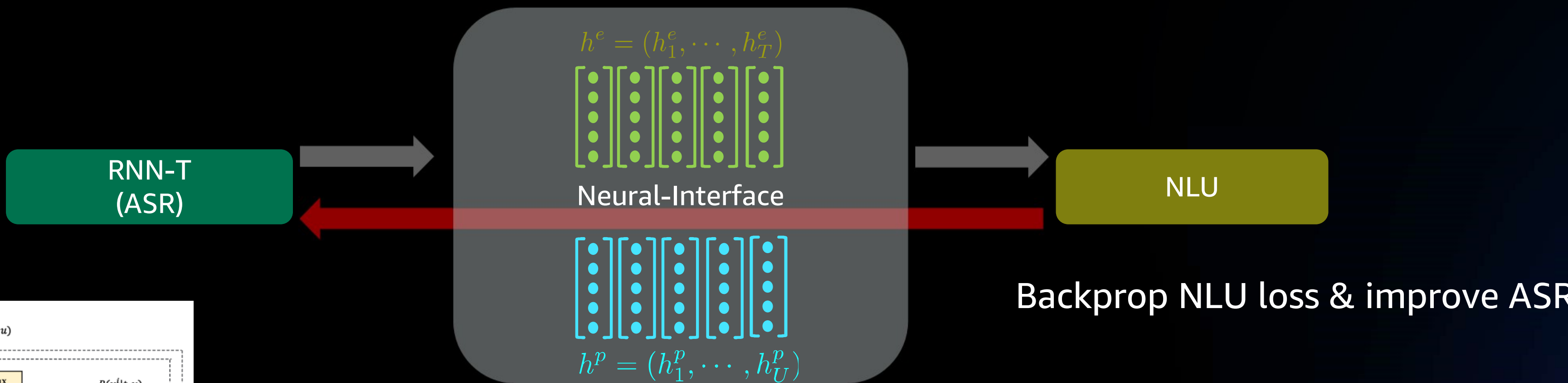


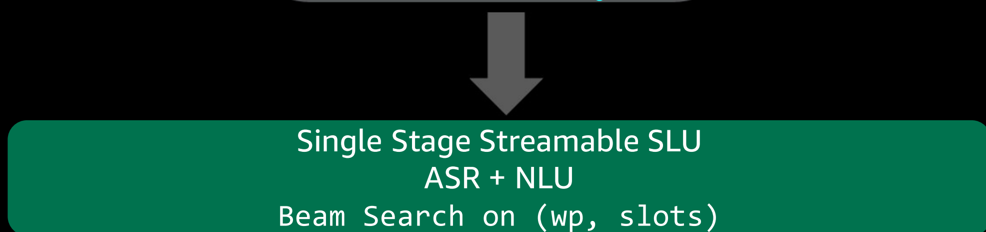
Backprop NLU loss & improve ASR

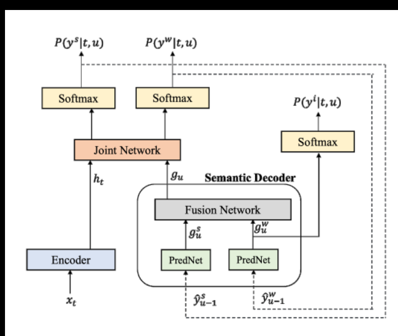
M. Rao, A. Raju, P. Dheram, B. Bui, A, Rastrow, "Speech to Semantics: Improve ASR and NLU Jointly via All-Neural Interfaces,", Interspeech 2020

A. Raju, G. Tiwari, et al., "End-to-end Spoken Language Understanding using RNN-Transducer ASR," arXiv preprint arXiv:2106.15919, 2021.

E2E Speech To Understanding







Model	Loss Type	WERR	SemERR	IRERR	ICERR
Two-stage SLU	-	0	0	0	0
Multi-task Semantic RNN-T	$L_{rnnt}(wp) + L_{ce}\ (slot) + L_{ce}\ (slot)$	**1.41**	**9.49**	**14.38**	**5.13**
	$L_{rnnt}(wp) + L_{rnnt,align}\ (slot) + L_{ce}\ (slot)$	-0.99	7.43	12.04	-1.26

X. Fu, F. Chang, M. Radfar, K. Wei, J. Liu, G. Strimel, K. M. Sathyendra, “Multitask RNN-T with Semantic Decoder for Streamable Spoken Language Understanding,” ICASSP 2022

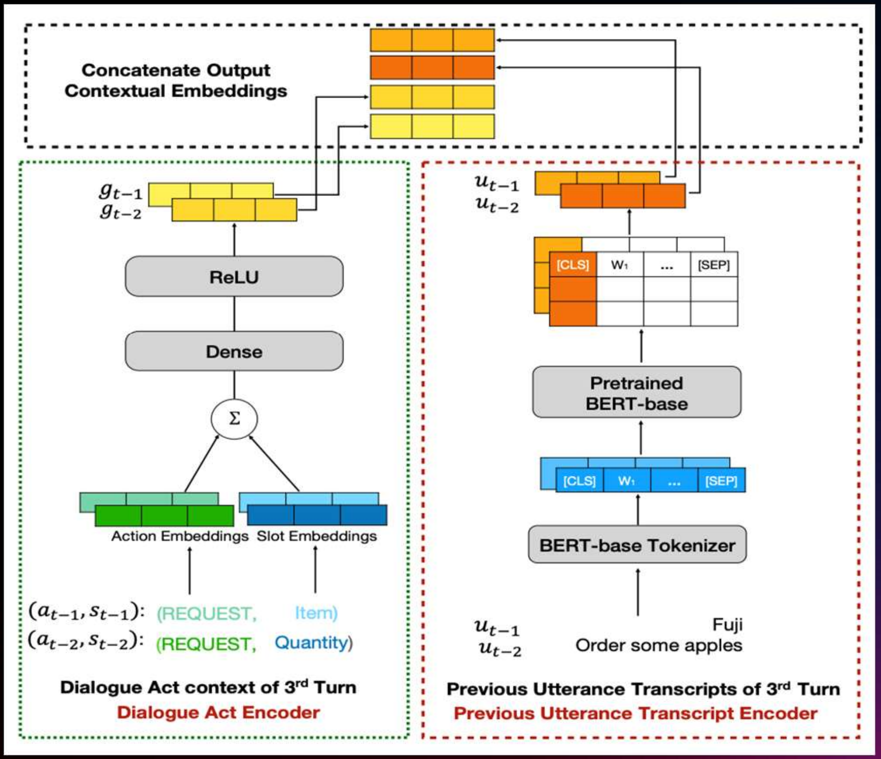
amazon | science

E2E SLU – Dialog Context Carry-Over

Transformer-based SLU w/ Context Carry-Over

- BERT embedding for transcription
- Multi-Head Attention with Gating for combining context
- Industrial Voice Assistant (IVA) Data Set

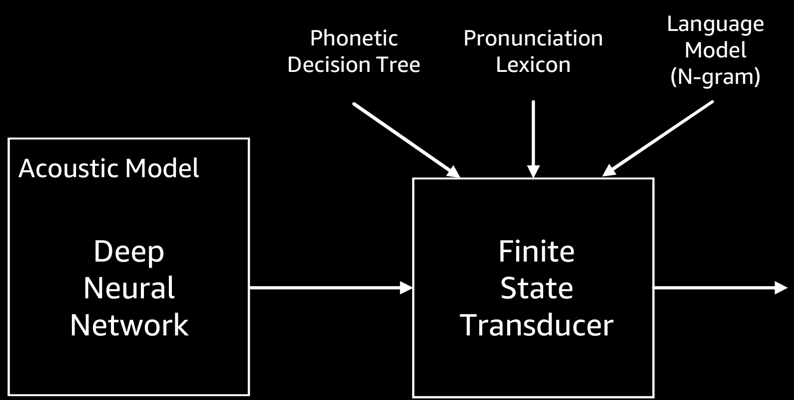
	Relative Error Reduction		
	WERR	ICERR	SemERR
E2E T-T SLU	0%	0%	0%
+ dialog act	5.4%	4.6%	1.5%
+ prev. utterance	12.4%	8.9%	6.3%
+ both	13.8%	11.1%	10.6%



K. Wei et al., "Attentive contextual carryover for multi-turn end-to-end spoken language understanding", ASRU 2021

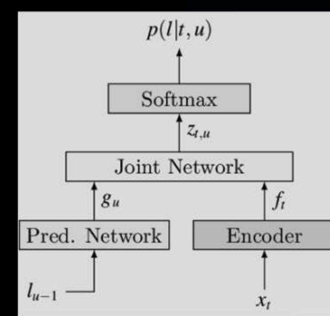
Edge Processing – Small Footprint ASR & SLU

Legacy factored HMM



100x compression

End-to-end all-neural



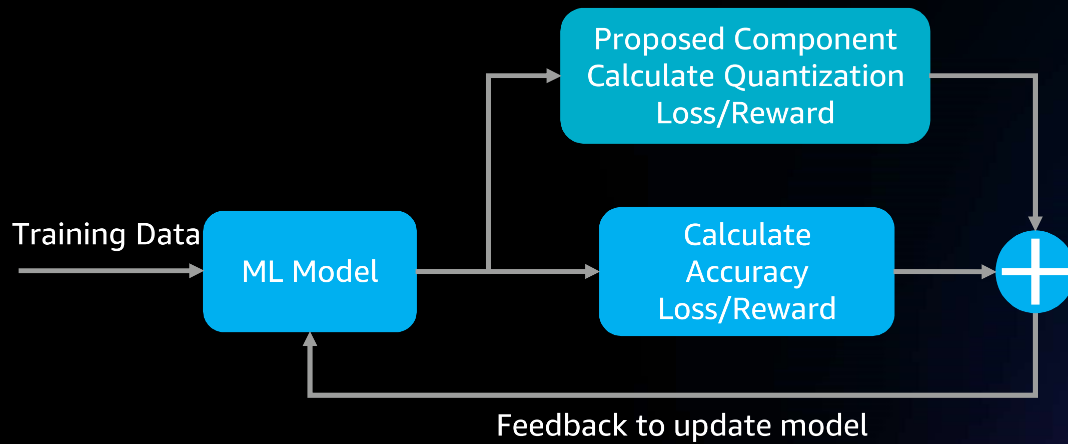
- N-grams are **memory inefficient**
- **Sub-optimal accuracy-vs-footprint** curve (disjoint models)

- Far better **accuracy-vs-footprint** curve
- Uniform application of **compression, quantization and sparsification** methods
  - 8-bit (and even 5-bit) quantization-aware training
- Architecture variation and choices
  - LSTM -> LSTM-P

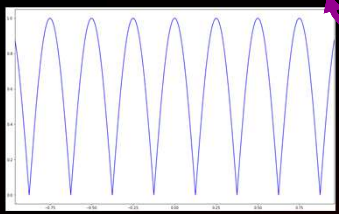
Edge Processing – Small Footprint ASR & SLU

Quantize-Aware Training via Regularization

Achieve 8-bit (and sub 8-bit)



Best weights = $\min\left[\mathcal{L}(weights) + \boxed{\mathcal{L}_{quantization}(weights)}\right]$

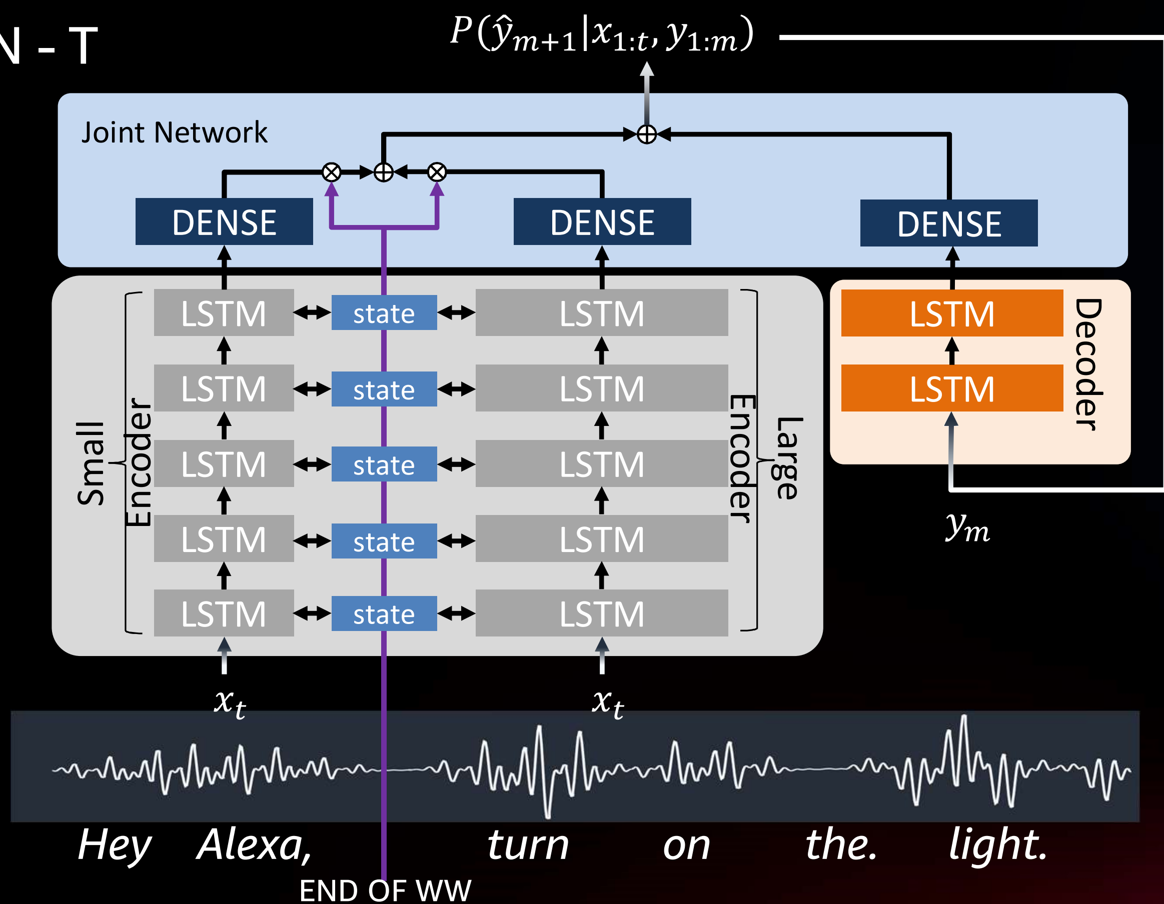


$\boxed{ACosR(x) = -\alpha|\cos(x)|}$

Hieu Nguyen et all, "Quantization aware training with absolute-cosine regularization for automatic speech recognition," Interspeech 2020

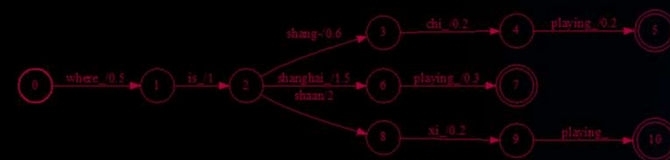
Edge Processing – Small Footprint ASR

Bifocal RNN-T



J. Macoskey et all, "Bifocal Neural ASR: Exploiting Keyword Spotting for Inference Optimization," ICASSP 2021

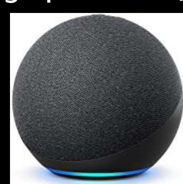
Edge Processing – Small Footprint ASR



Neural Transducer (RNN-T) → Generate N-best list/Lattice → Neural Rescorer (text + lattice) → Generate N-best list/Lattice → Neural LM Rescorer (text)

On-device
(using Neural Edge processor)

In-Cloud
(Unified Rescoring Pipeline with Cloud-based Devices)



Conclusions

- What we have briefly touched
  - Dynamic Adaptation and Personalization
    - Attention-based Neural Biasing
  - E2E Speech To Understanding
    - Backpropagate NLU loss & improve ASR
    - Semantic decoder & fusion network
    - Dialog Context Carry-Over
  - Small Footprint ASR
    - Quantization aware training
    - Bi-focal RNN-T

What we haven't covered

- Representation Learning
- Multi-Lingual Modeling
- Multi-Speaker Modeling
- Multi-Modal Modeling
- Closed-loop self-learning,
  Semi-/weakly-supervised learning
- Life-long learning
- Learning on device
- …

It is still Day One!

A good time to be a speech researcher!

Thank you!

Jimmy Kunzmann
kunzman@amazon.com

Bjorn Hoffmeister
bjornh@amazon.com

Ariya Rastrow
arastrow@amazon.com

Daniel Willett
dawillet@amazon.de