New Frontiers for Speech and Language Processing

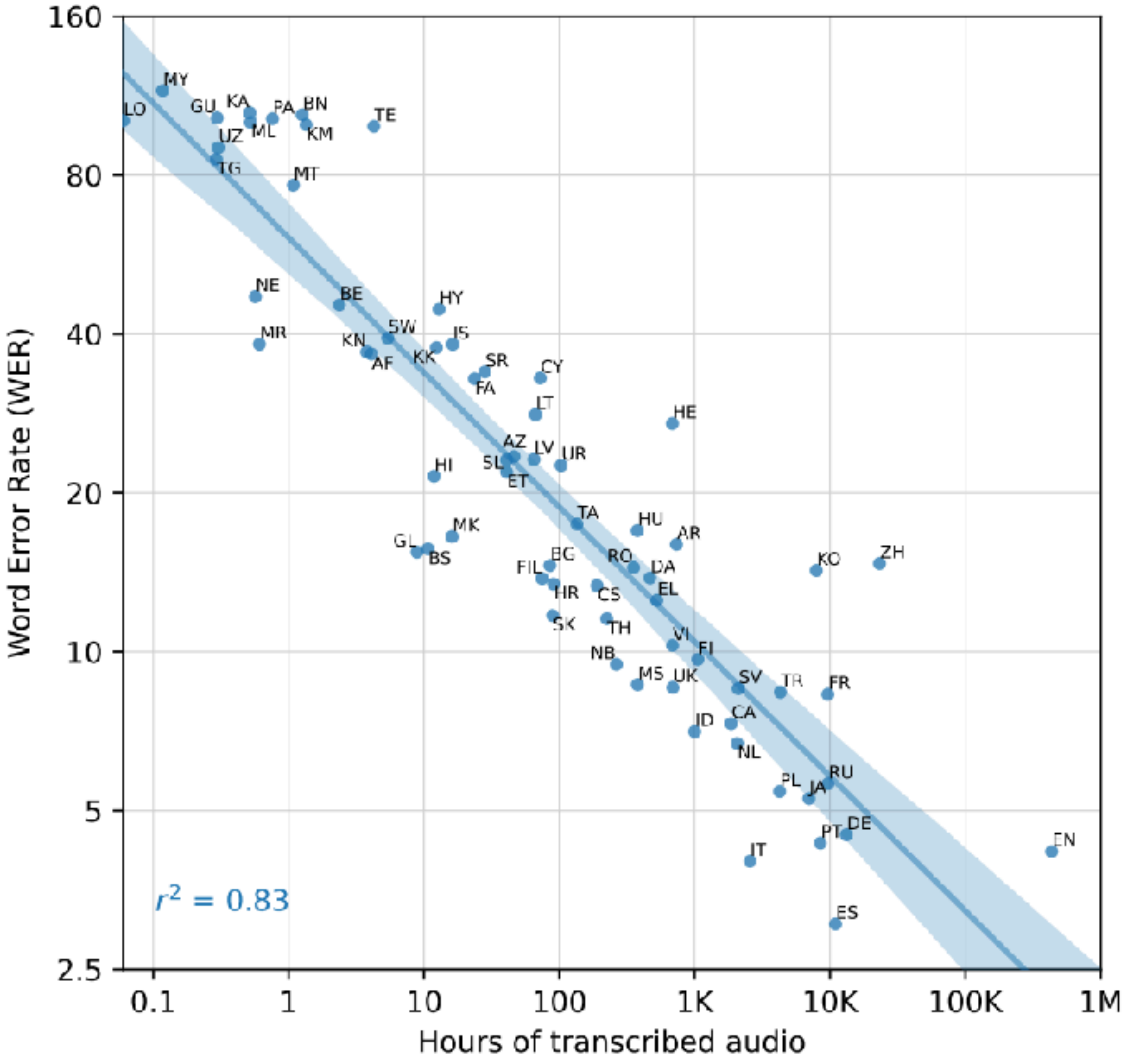
An Indian Language Perspective

Preethi Jyothi, IIT Bombay
CS Katha Barta, NISER
March 28, 2023

Speech and Language Technologies (SLT) for India

- SLT faces significant challenges in India
    - With hundreds of languages, thousands of dialects*
- High correlation between supervision for a language/accent and its final WER [1]

> "We observe lower accuracy on low-resource and/or low-discoverability languages or languages where we have less training data. The models also exhibit disparate performance on different accents and dialects of particular languages."
>
> https://github.com/openai/whisper/blob/main/model-card.md

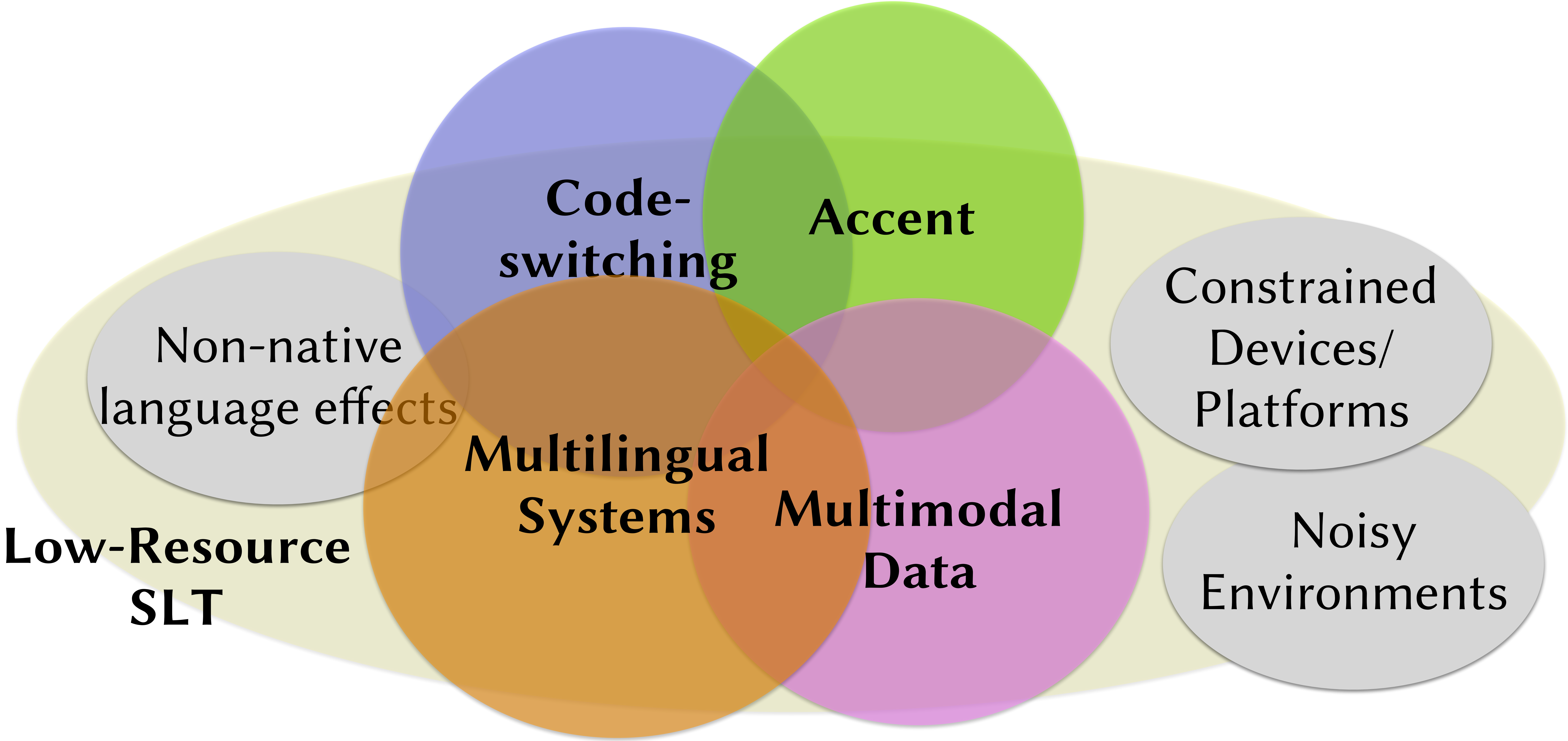


* Census 2011: 19,569 raw linguistic affiliations, 1369 rationalized mother tongues

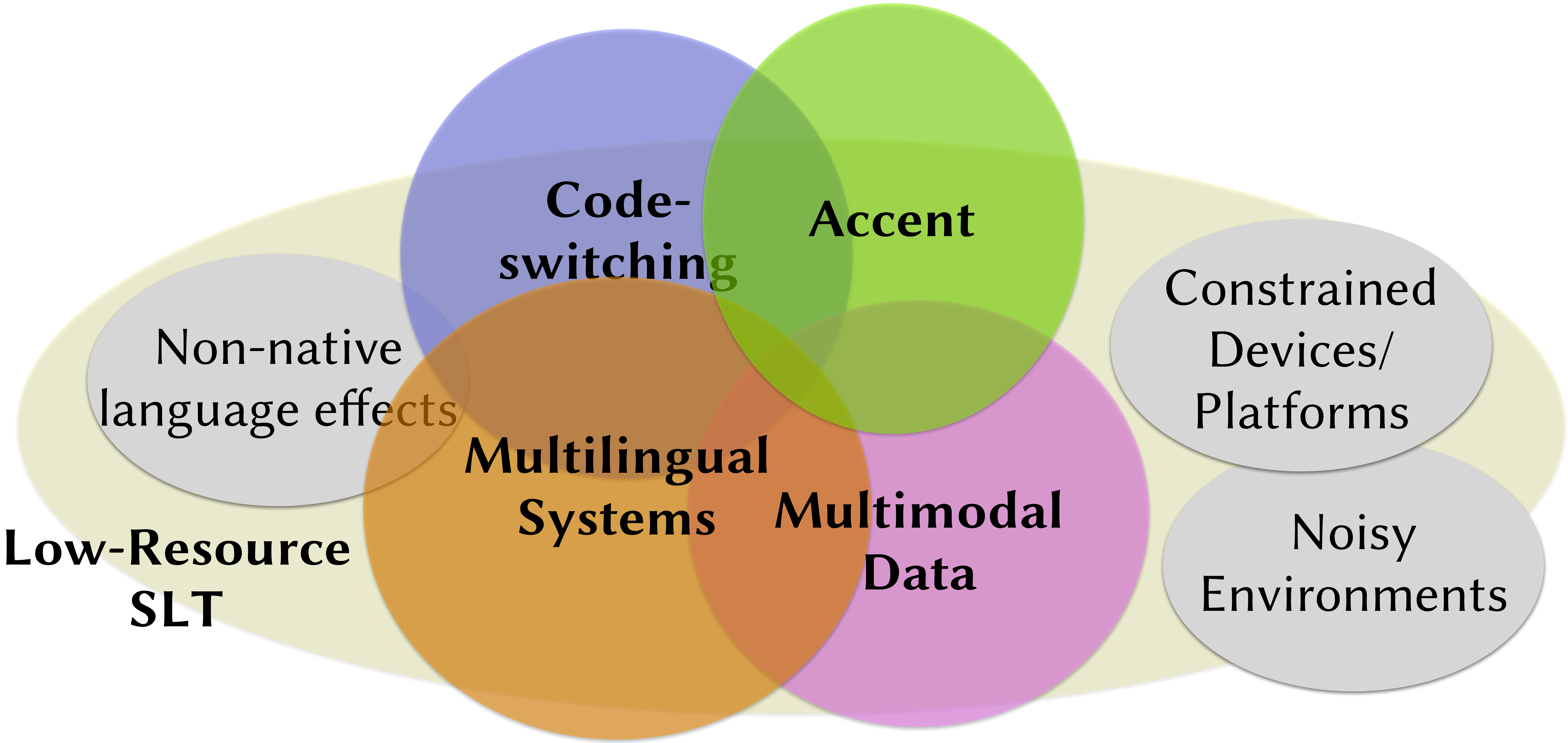
[1] "Robust Speech Recognition via Large-scale Weak Supervision", Radford et al., https://arxiv.org/pdf/2212.04356.pdf, Dec 2022

Speech and Language Technologies (SLT) for India

- SLT faces significant challenges in India
    - With hundreds of languages, thousands of dialects*
- High correlation between supervision for a language/accent and its final WER [1]

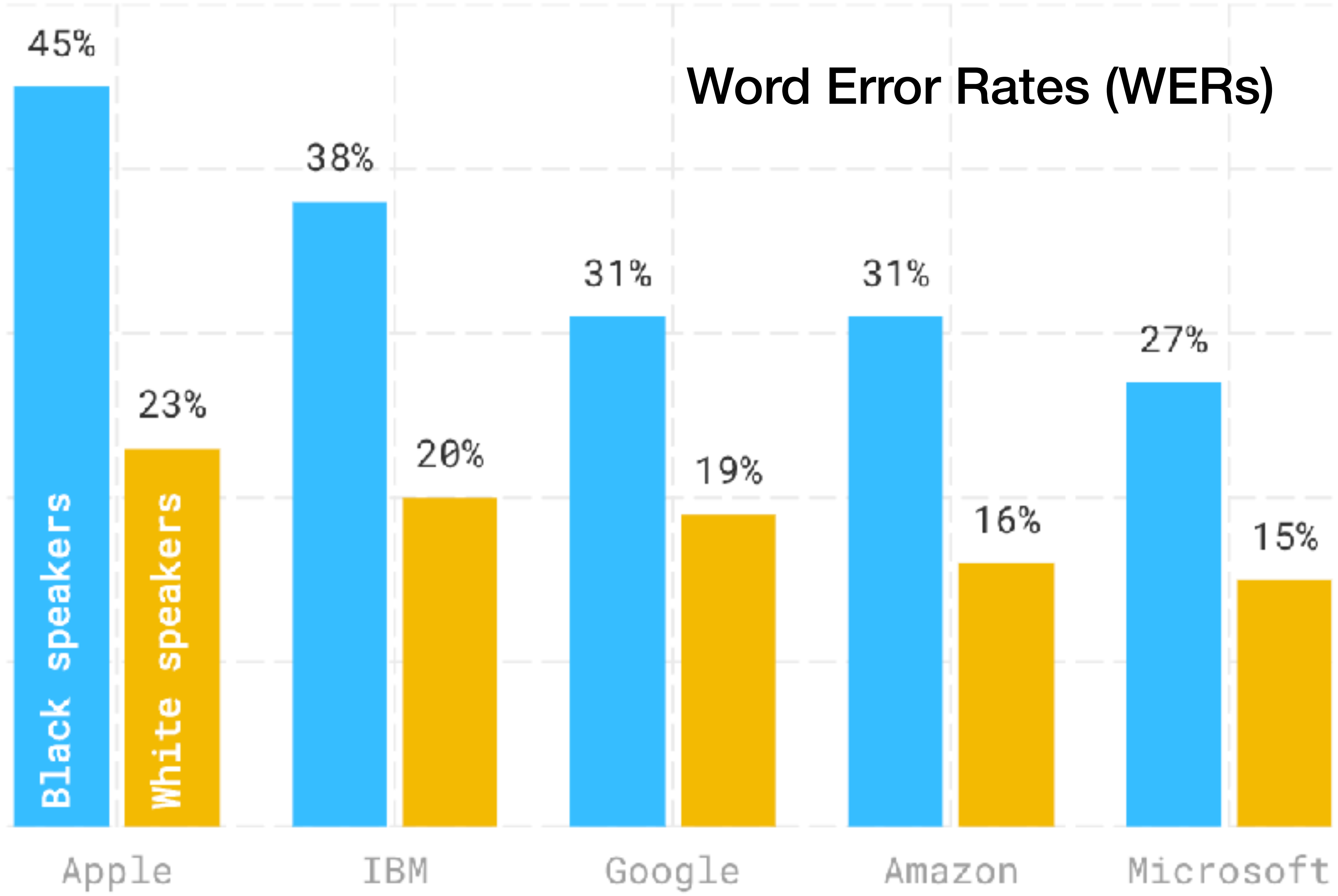


Speech and Language Technologies (SLT) for India



Voice Is the Next Big Platform, Unless You Have an Accent

- Non-native accents pose a significant challenge to state-of-the-art ASR systems



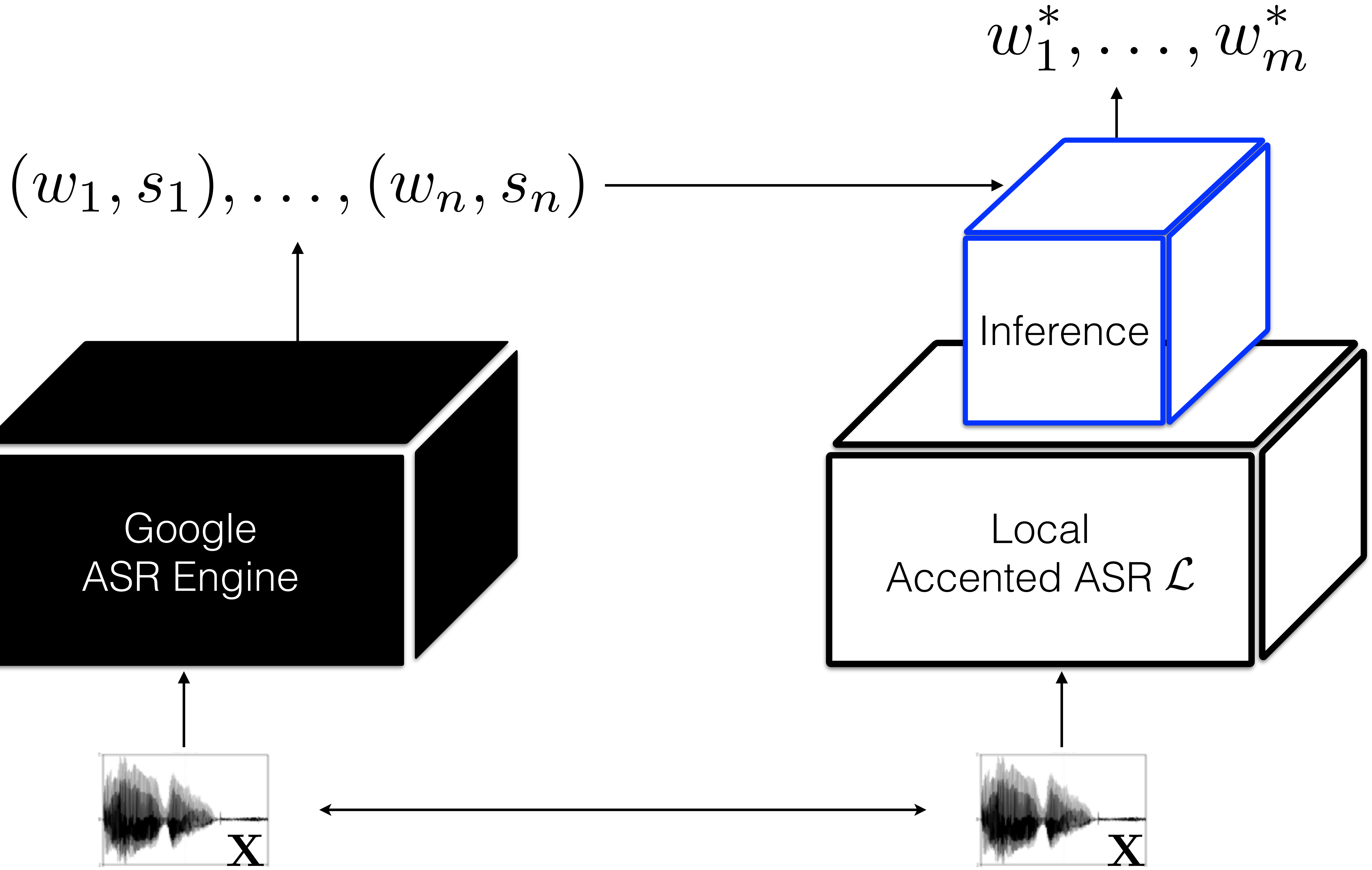
- Can we use blackbox service APIs to guide a local ASR system targeting specific accents?

Image from https://fairspeech.stanford.edu/, 2020

Adapting Black-box ASR Systems to Accented Speech

- Guided inference to adapt a black-box ASR system to speech from a target accent

KJAS'20



KJAS'20 K. Khandelwal, P. Jyothi, A. Awasthi and S. Sarawagi, Black-box Adaptation of ASR for Accented Speech, Interspeech 2020

Adapting Black-box ASR systems to Accented Speech

- Guided inference to adapt a black-box ASR system to speech from a target accent

- We propose a guided inference algorithm (***FineMerge***) KJAS'20
  - Build a local ASR system L specific to the target accent
  - Predicts character distributions $P_1, \ldots, P_T \triangleq \mathbf{P}$ for T input frames at test time
  - Align service characters from $\mathbf{s}$ to each frame using $\mathbf{P}$ to get $S_1, \ldots, S_T$
  - Revise $\mathbf{P} \rightarrow \mathbf{P}^s$ to selectively support service characters

S_t	_	p	o	_	_	s	t	e	d	d
$P_t(S_t)$	6e-5	1e-11	1	0.34	0.01	0.93	0.99	0.44	0.29	0.98
d_t	t	t	o	o		s	t	a	t	d
$P_t(d_t)$	0.99	0.99	1.0	0.63	0.98	0.93	0.99	0.55	0.64	0.98
r_t	t	t	o	_		s	t	e	d	d
$P_t^s(r_t)$	0.62	0.99	1.0	0.59	0.61	0.93	0.99	0.66	0.57	0.98
$P_t(r_t)$	0.99	0.99	1.0	0.34	0.98	0.93	0.99	0.44	0.29	0.98

KJAS'20 K. Khandelwal, P. Jyothi, A. Awasthi and S. Sarawagi, Black-box Adaptation of ASR for Accented Speech, Interspeech 2020

Black-box ASR Adaptation

Method	WER (Indian En)	WER (Australian En)	WER (British En)
Local	27.99	24.41	25.06
Service	22.32	23.52	20.82
Rover	21.12	18.04	18.10
FineMerge	**18.45**	**16.90**	**16.47**

[1] "Black-box Adaptation of ASR for Accented Speech", K. Khandelwal, P. Jyothi, A. Awasthi, S. Sarawagi, Interspeech 2020

Black-box ASR Adaptation

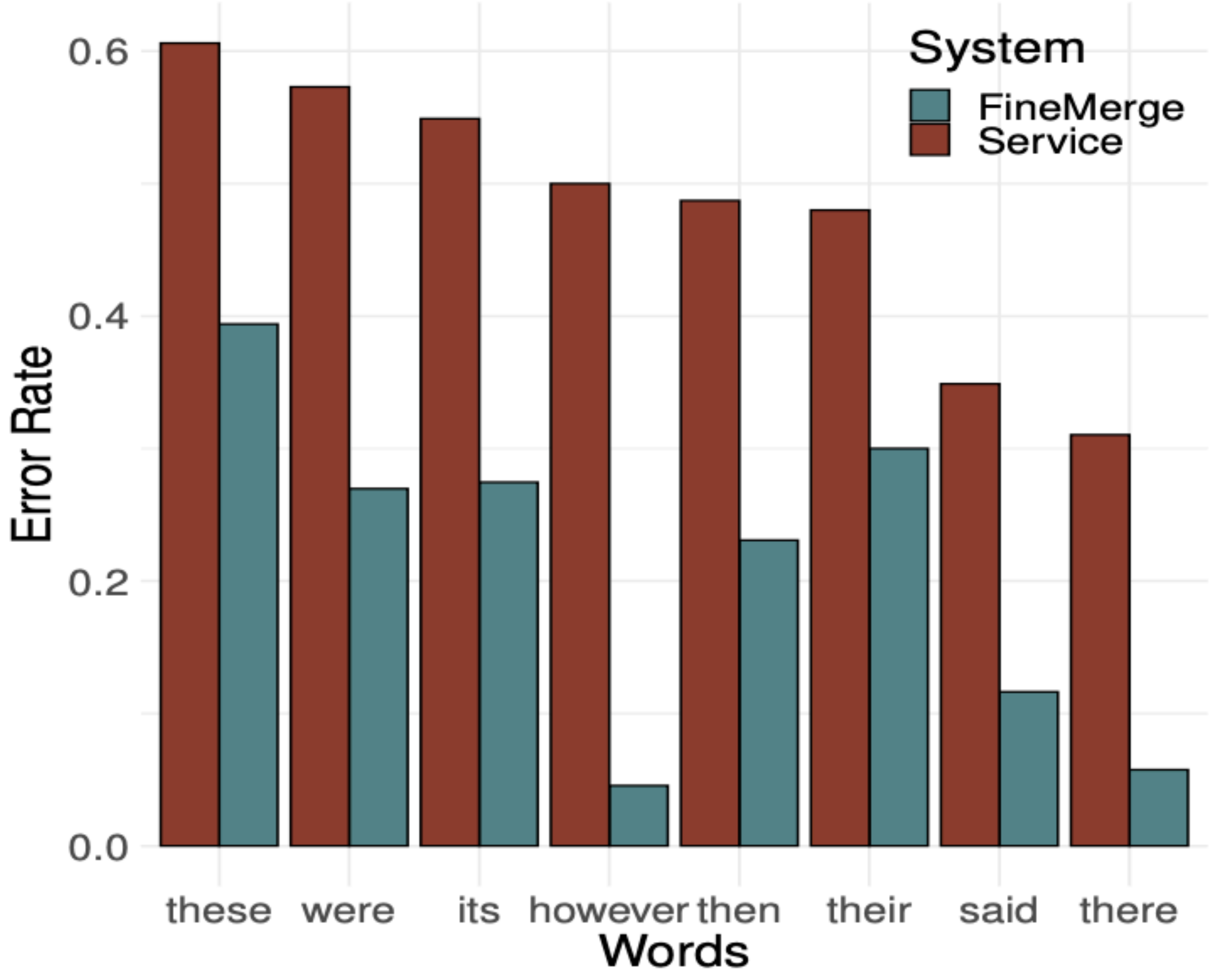
Method	WER (Indian En)	WER (Australian En)	WER (British En)
Local	27.99	24.41	25.06
Service	22.32	23.52	20.82
Rover	21.12	18.04	18.10
FineMerge	**18.45**	**16.90**	**16.47**

	Indian	Australian
Gold	for a brief time …	… rope a bull while on a
Service	soda beef time …	… work a bowl while on a
Local	for a breeze time …	… rope the ball while on a
Rover	for a beef time …	… work a bowl while on a
FineMerge	for a **brief** time …	… rope a **bull** while on a

[1] "Black-box Adaptation of ASR for Accented Speech", K. Khandelwal, P. Jyothi, A. Awasthi, S. Sarawagi, Interspeech 2020

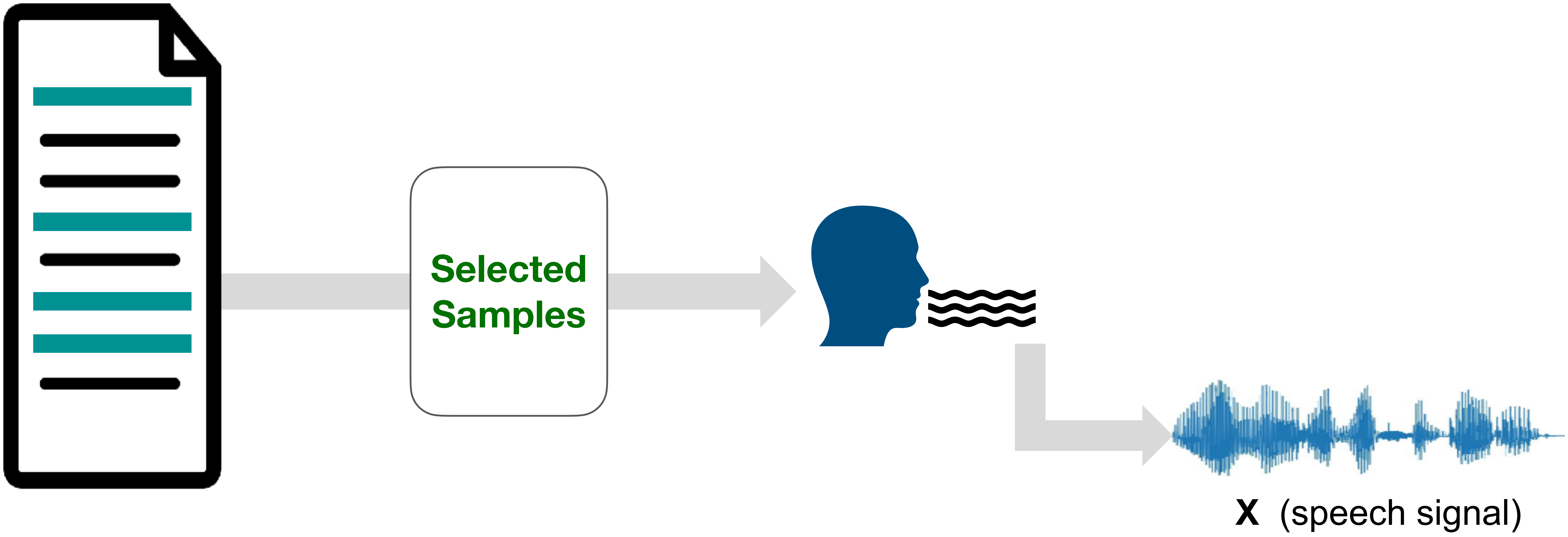
Black-box ASR Adaptation

- Words with highest reductions in error on Indian-accented test samples
- "however": Contains the diphthong /aw/ that has many phonetic realizations
- "were": /v/ and /w/ usually overlap in usage by Indian-accented English speakers



Personalization: Accent Adaptation For a Specific Speaker

- For personalised ASR, collect speech by asking users to read out selected samples
- How do we select samples? Can we do better than random selection?



Sentence Selection

Pick examples that are more (ASR) error-prone AKSJ'21

Finding sentences that are ASR error-prone:

1. Learn an **"error model"** that identifies phonemes in a sentence that ASR may misrecognize
2. Use a small seed set to train the error model
3. Assign higher scores to sentences with more errors using the error model

AKSJ'21 A. Awasthi, A. Kansal, S. Sarawagi, P. Jyothi Error-driven Fixed-budget ASR Personalization for Accented Speakers, ICASSP 2021

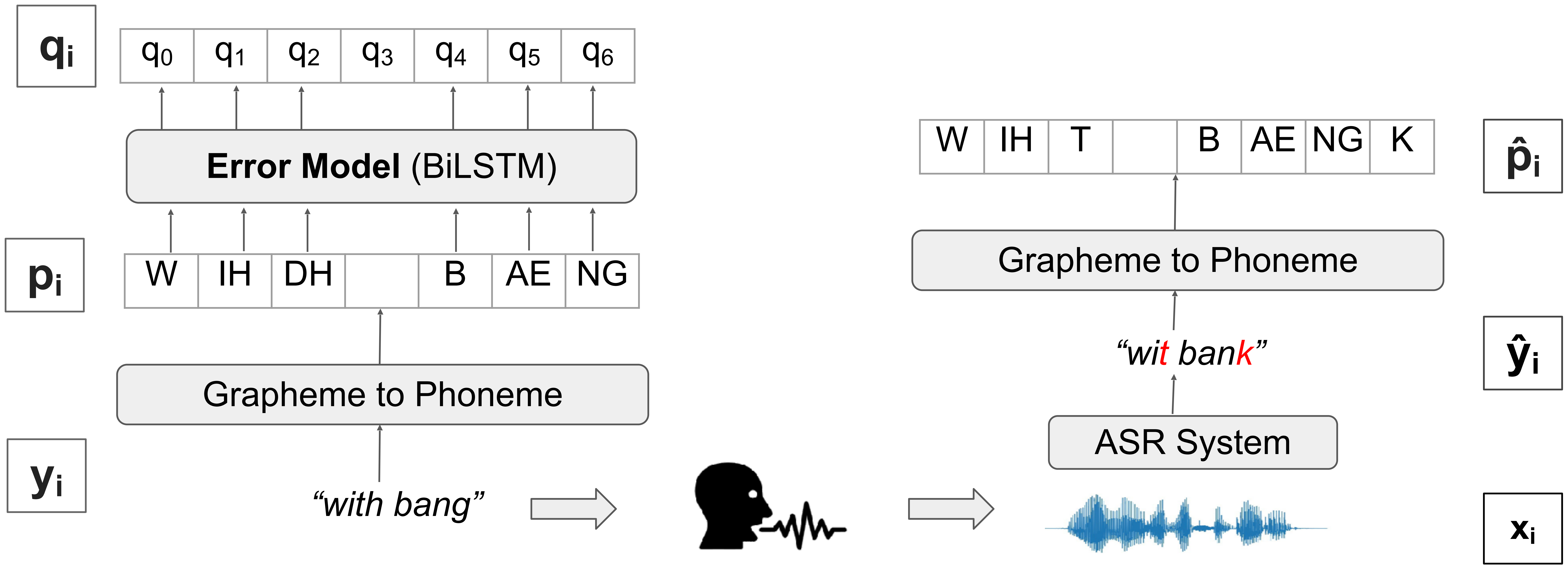
Training the Error Model

$\boxed{y_i}$

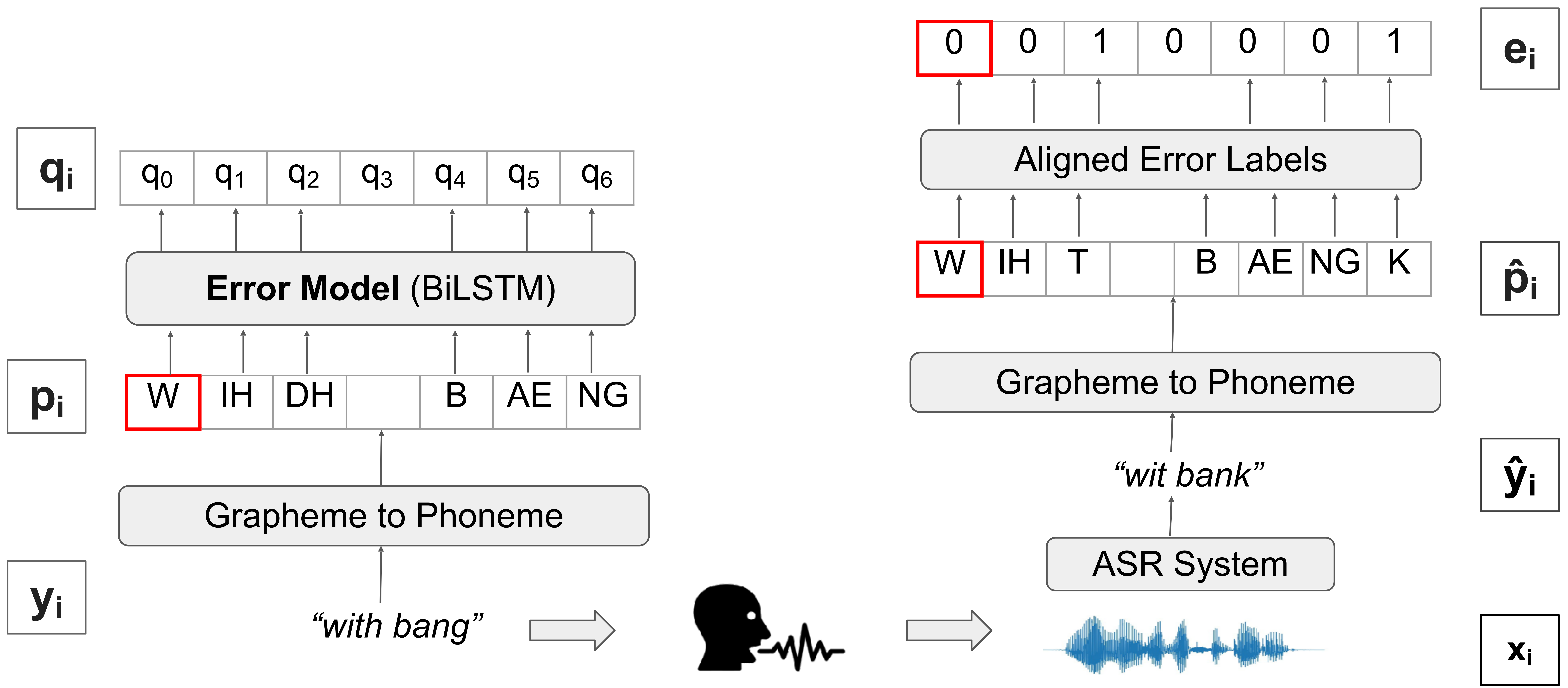
"with bang"

Next few slides made by Abhijeet Awasthi

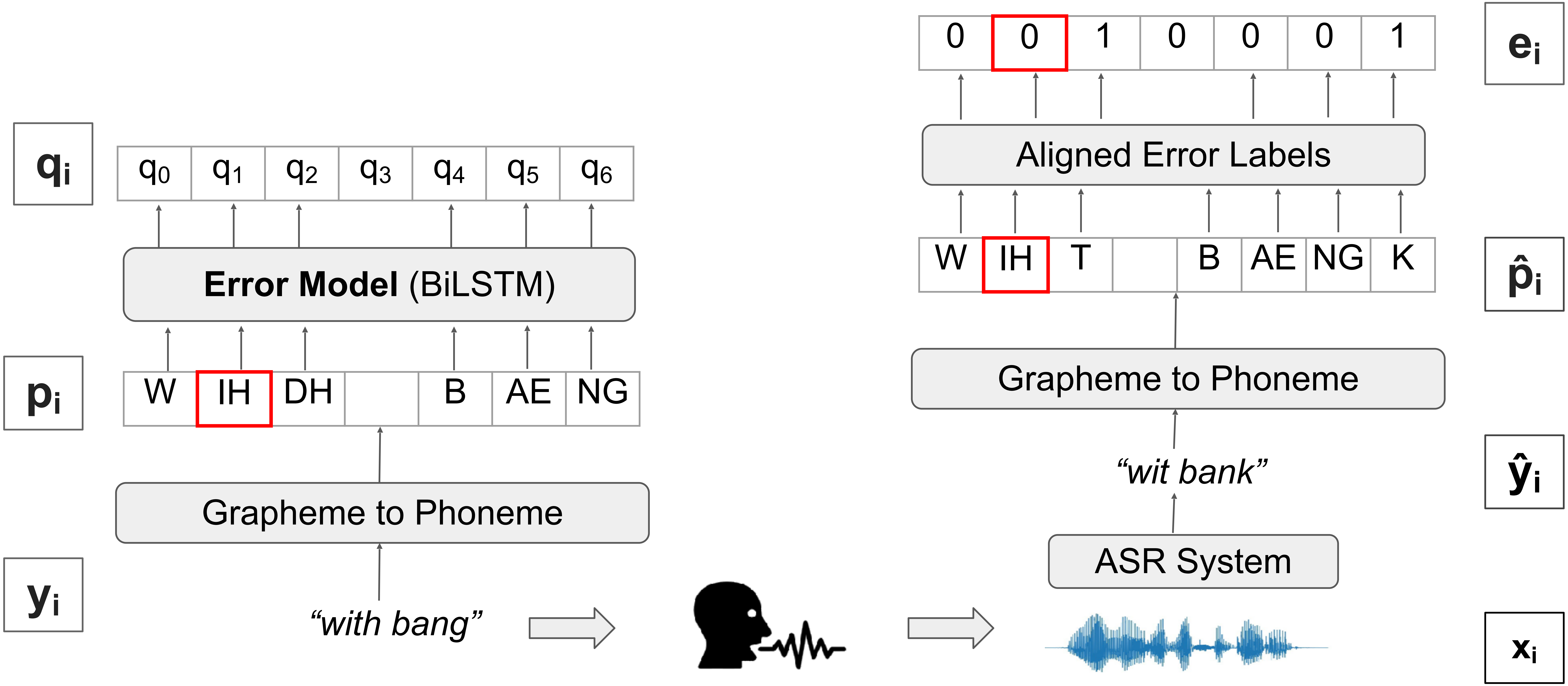
Training the Error Model



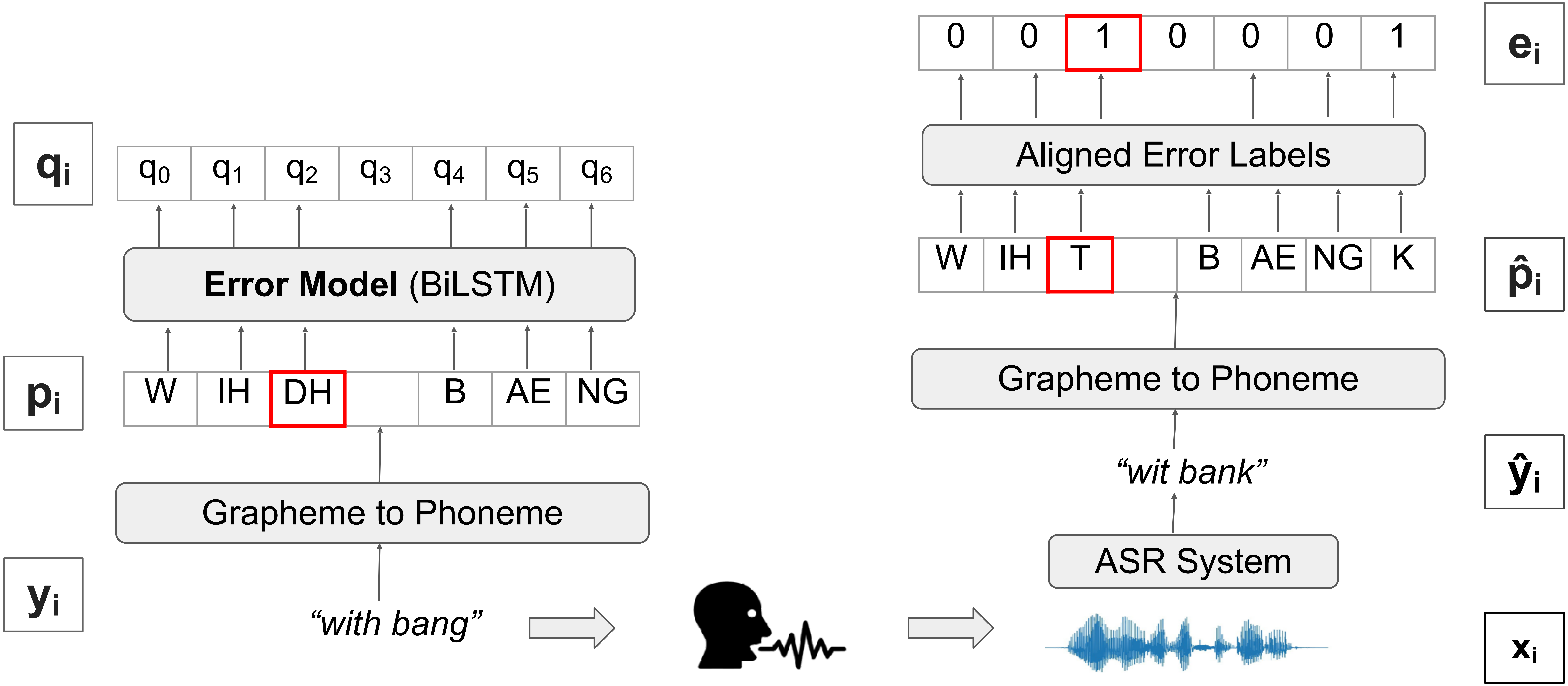
Training the Error Model



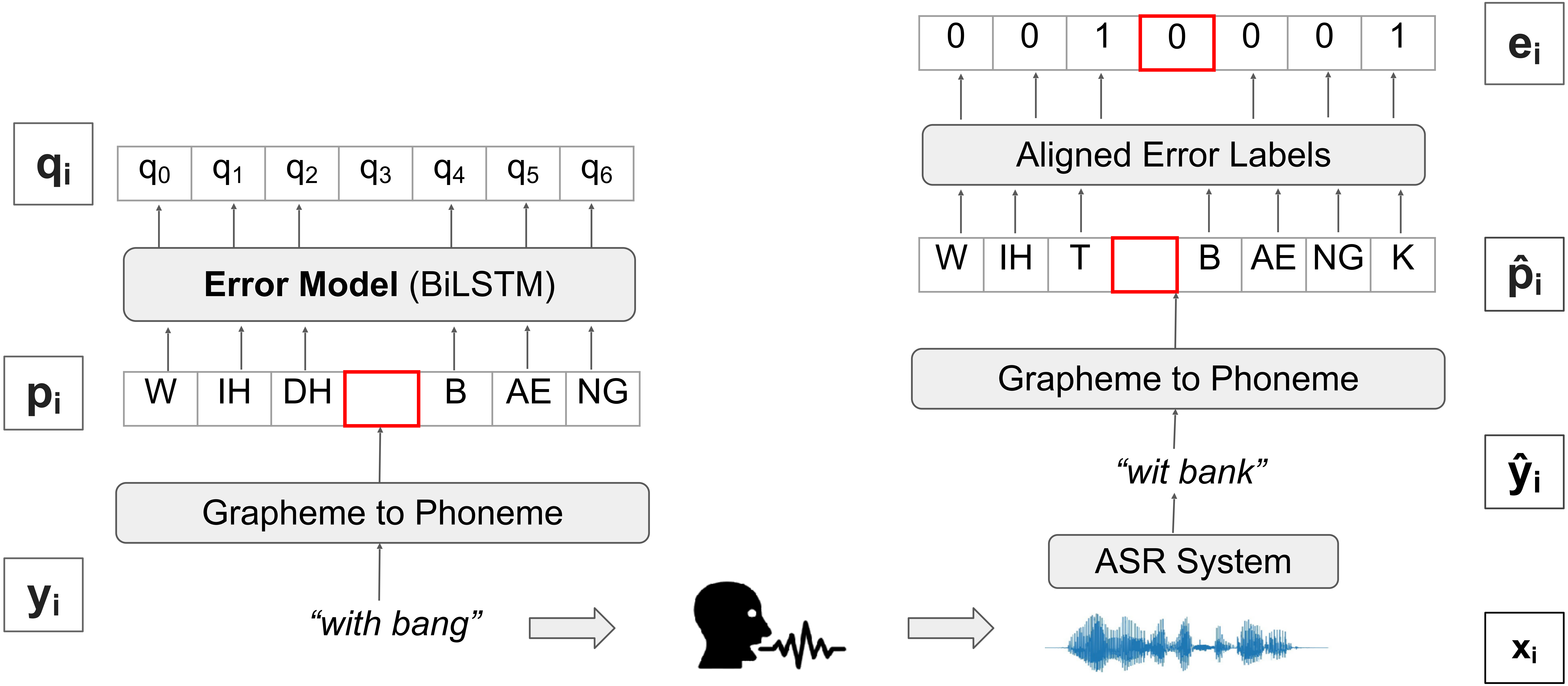
Training the Error Model



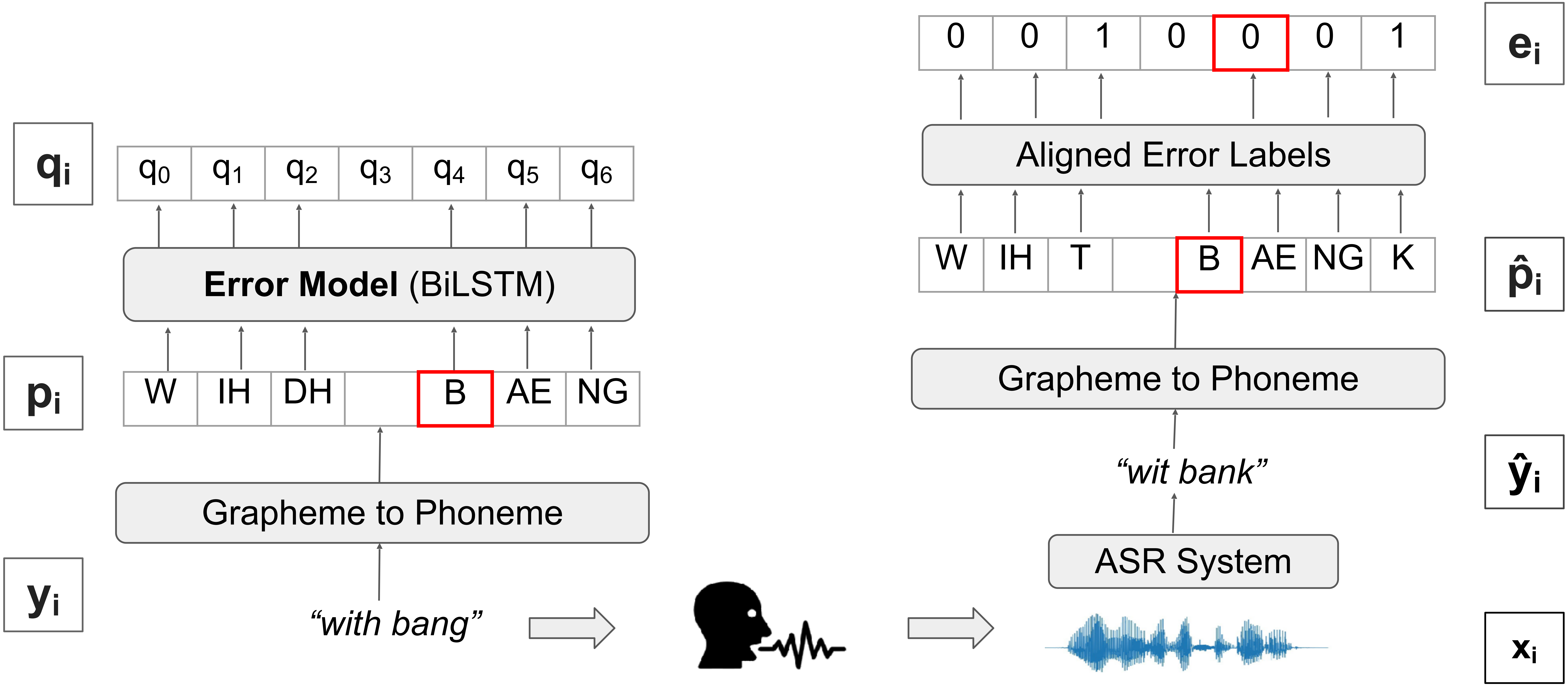
Training the Error Model



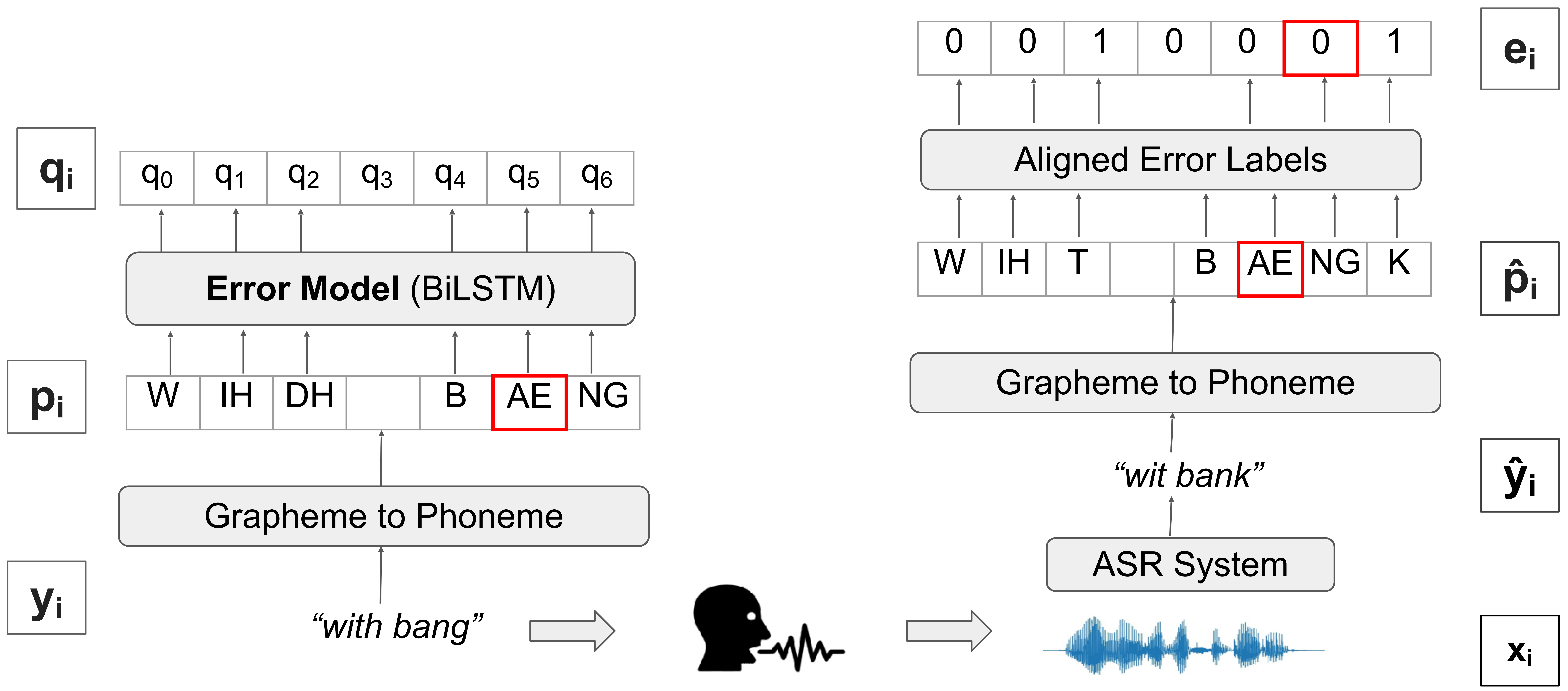
Training the Error Model



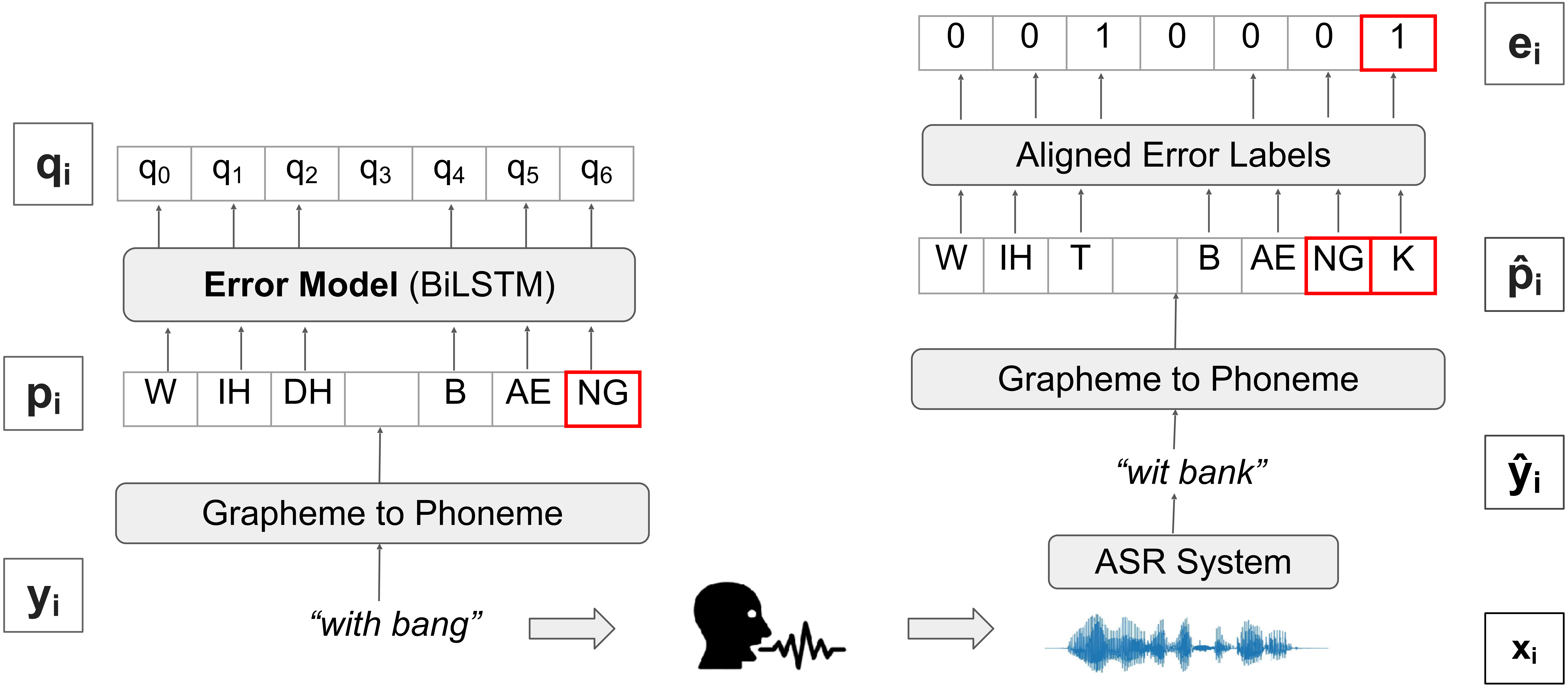
Training the Error Model



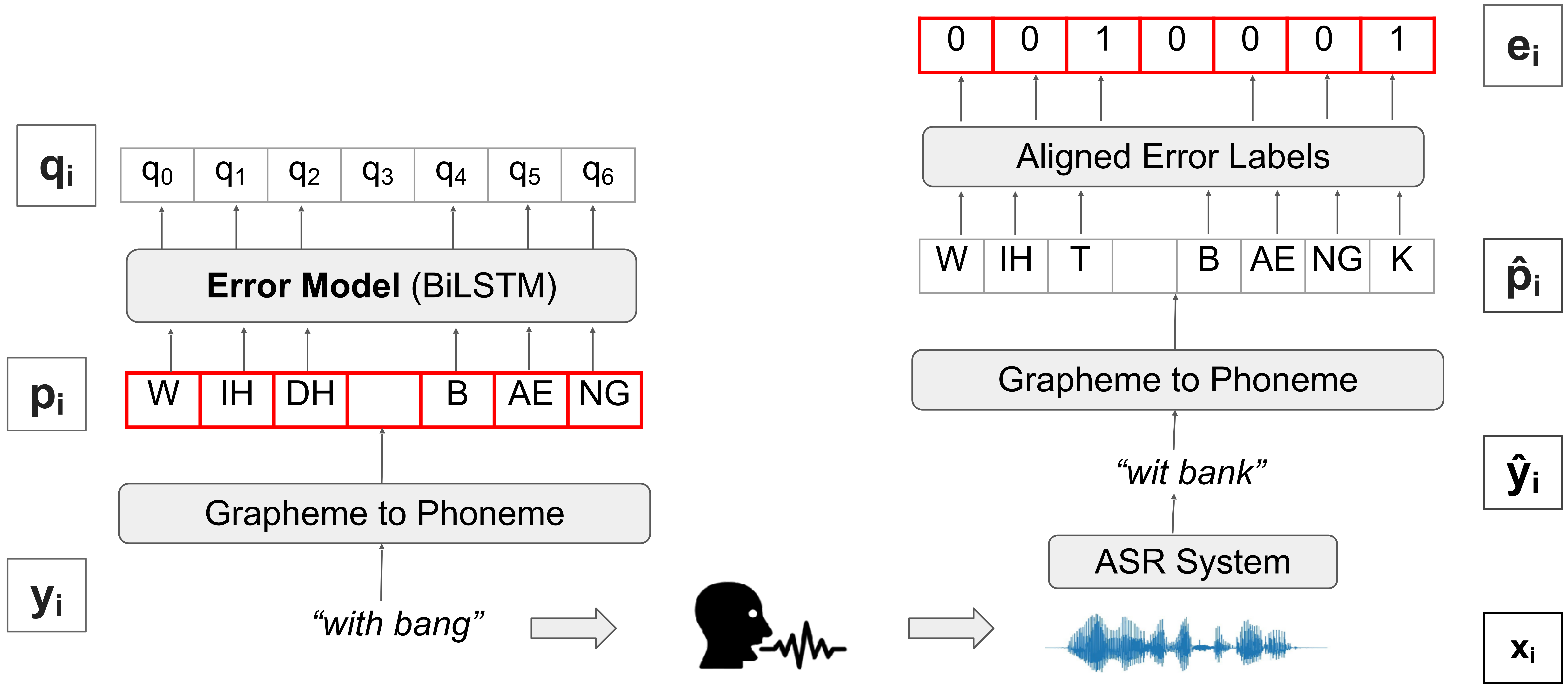
Training the Error Model



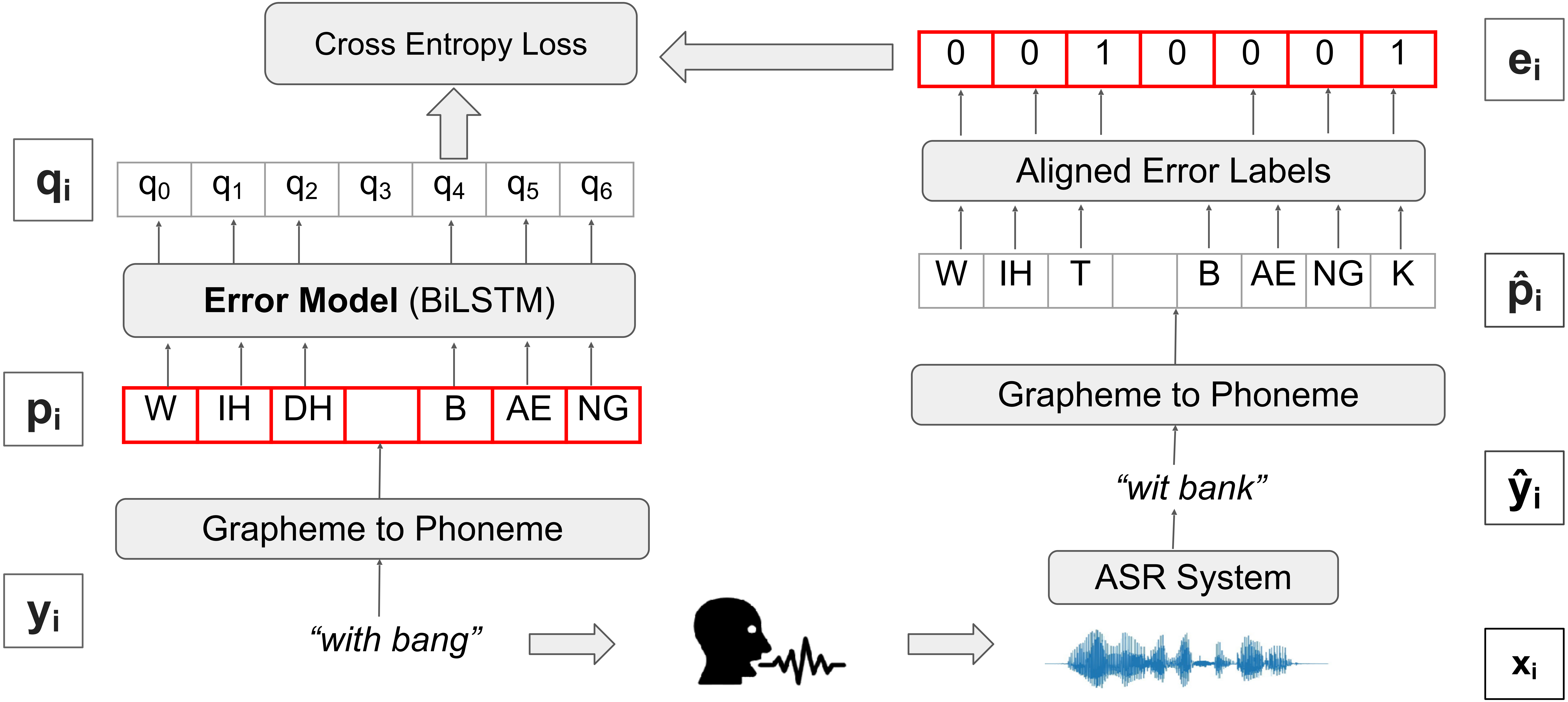
Training the Error Model



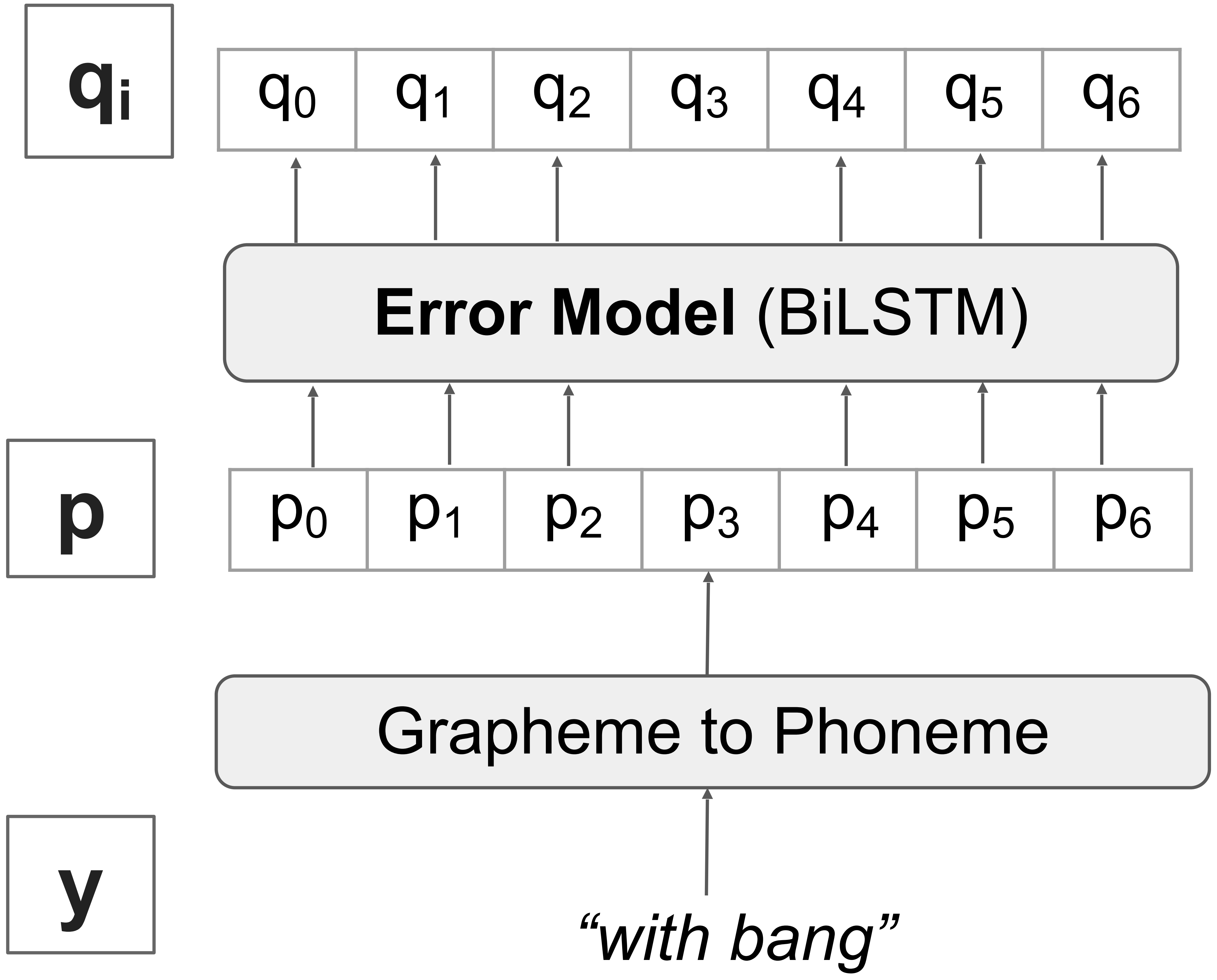
Training the Error Model



Training the Error Model

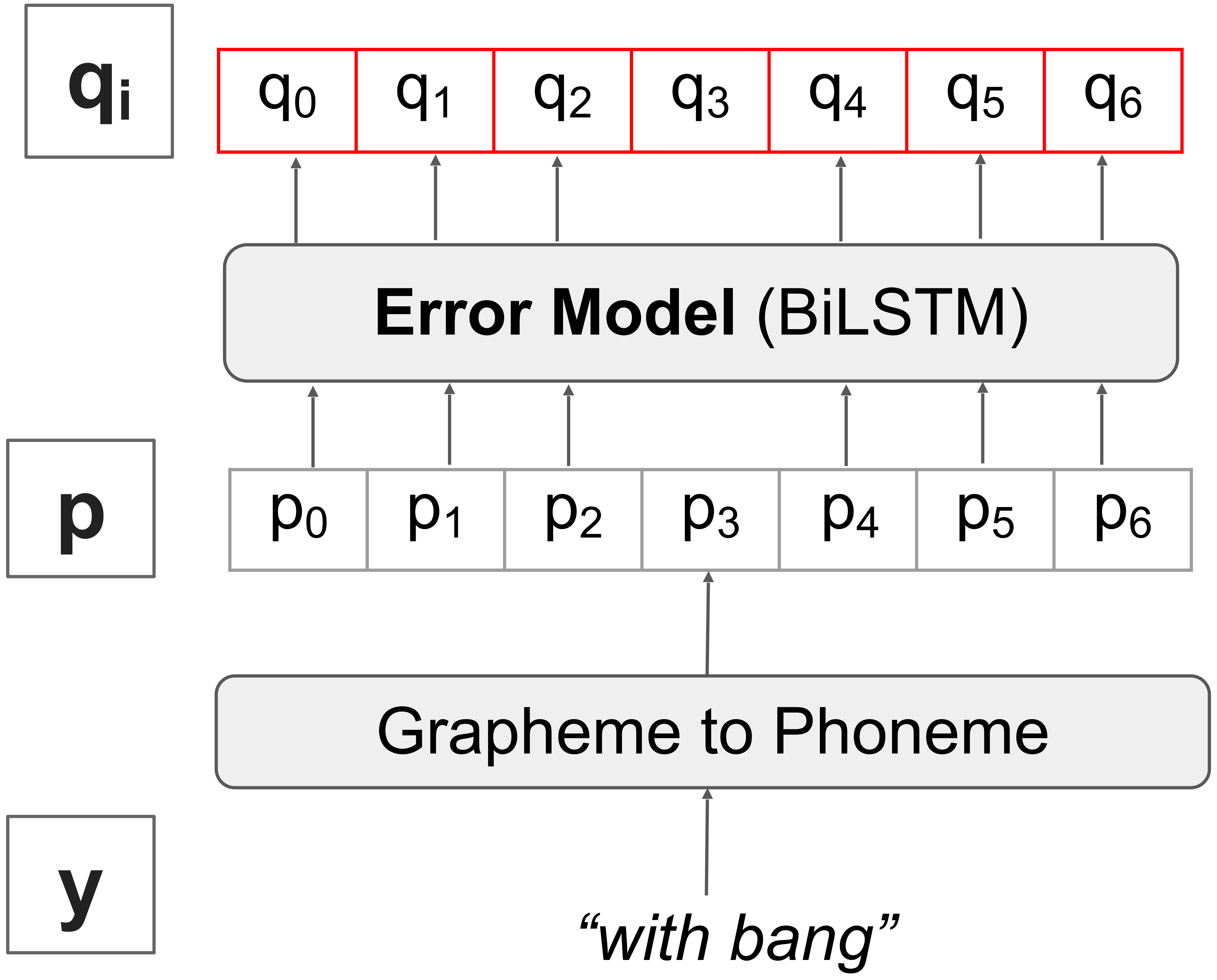


Scoring and Selecting Utterances using Error Model



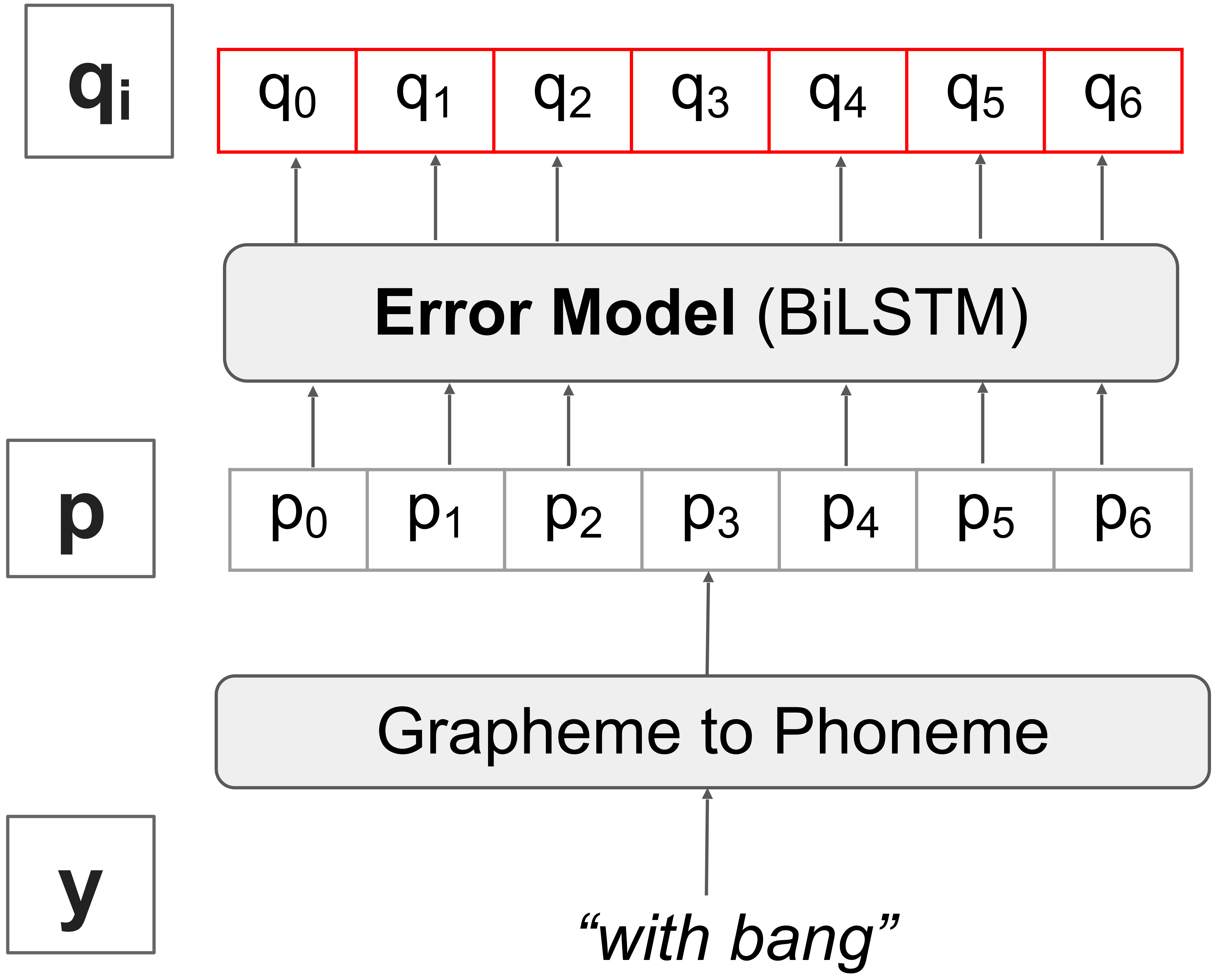
$$\text{score}(\mathbf{y}) = \frac{1}{n}\sum_{j} q_j$$

Scoring and Selecting Utterances using Error Model



$$\text{score}(\mathbf{y}) = \frac{1}{n} \boxed{\sum_j q_j}$$

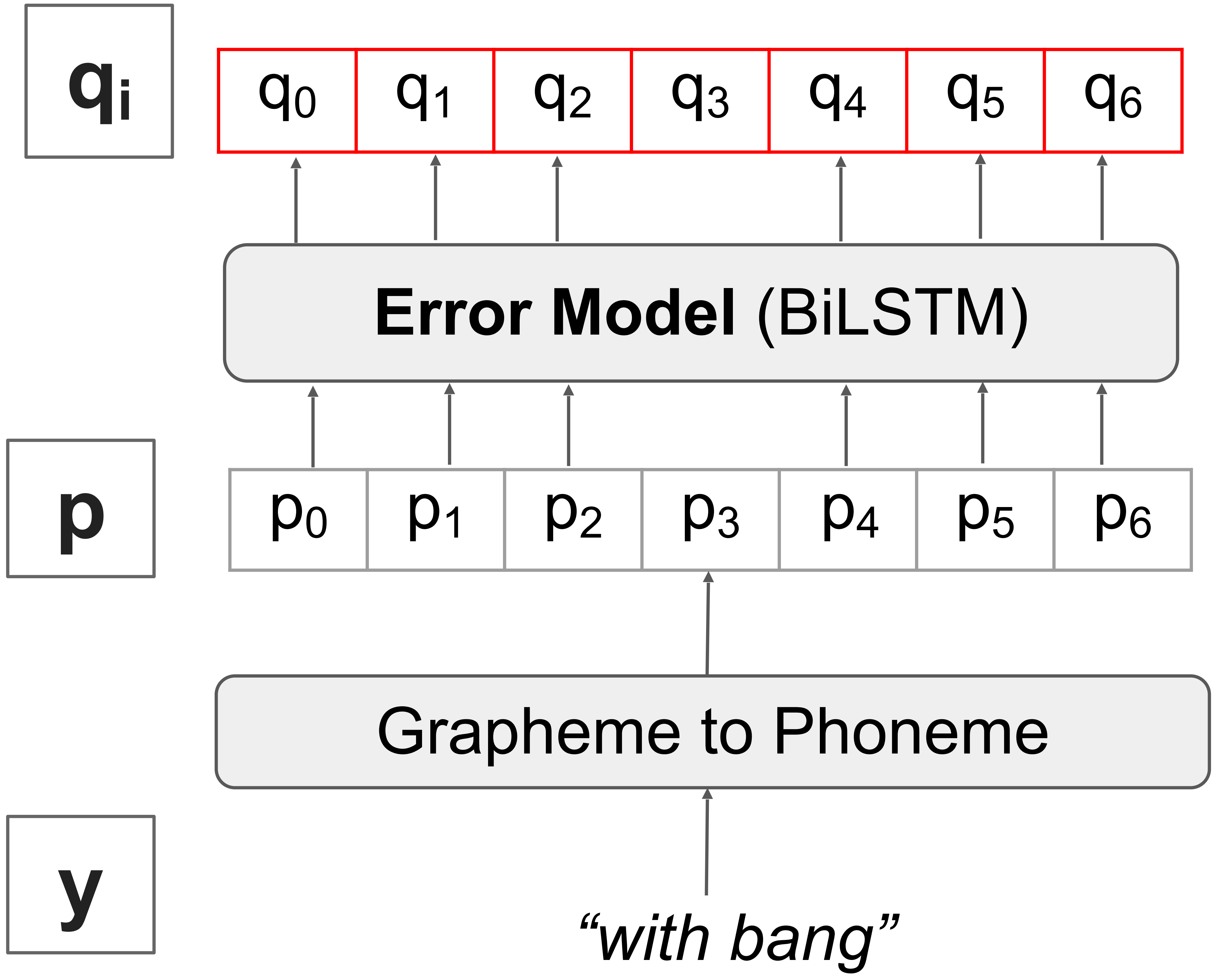
Scoring and Selecting Utterances using Error Model



$$\mathrm{score}(\mathbf{y}) = \frac{1}{n}\boxed{\sum_j q_j}$$

Prevents bias towards longer sentences

Scoring and Selecting Utterances using Error Model

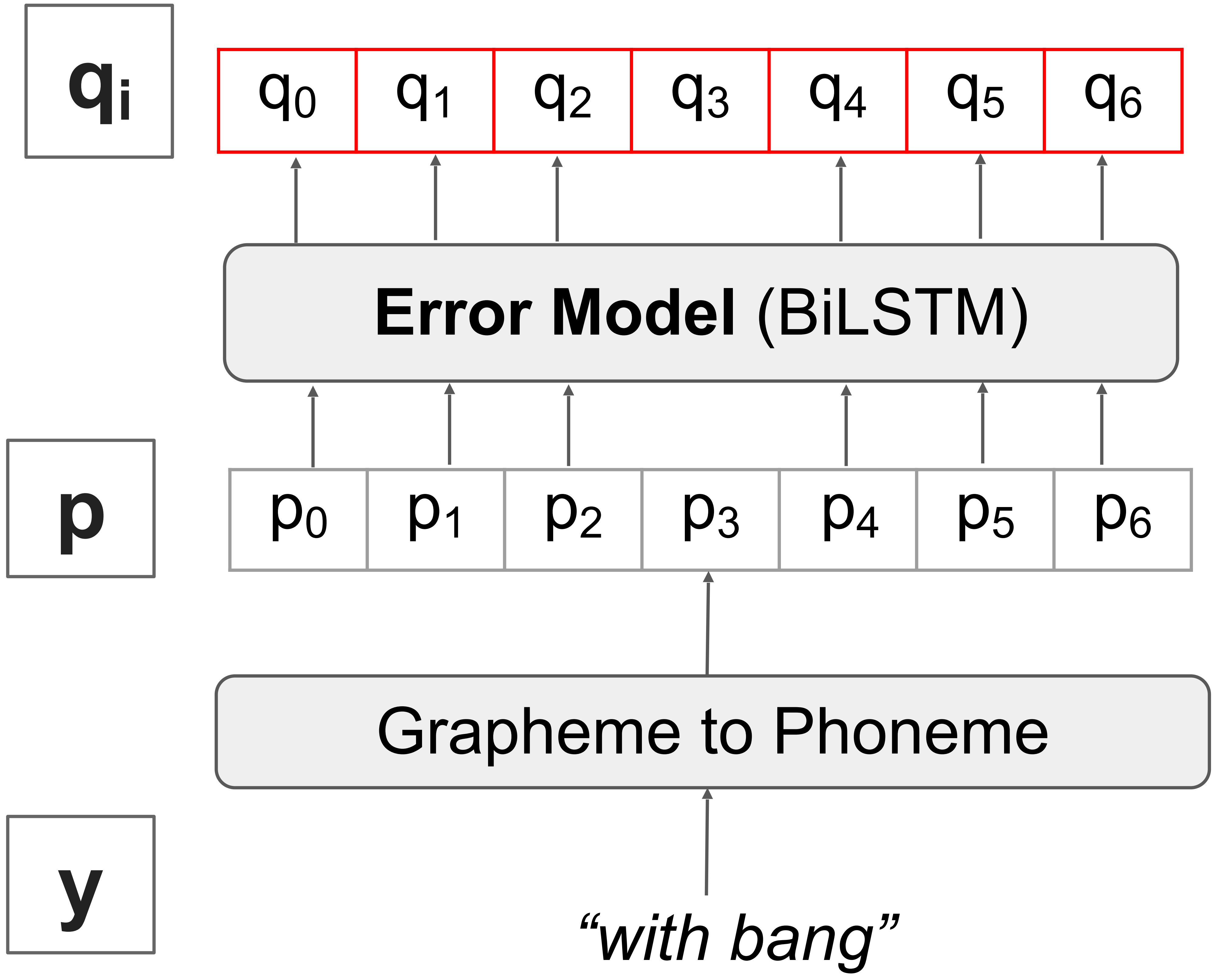


Reduces diversity by selecting repetitive patterns

$$\mathrm{score}(\mathbf{y}) = \frac{1}{n} \boxed{\sum_j q_j}$$

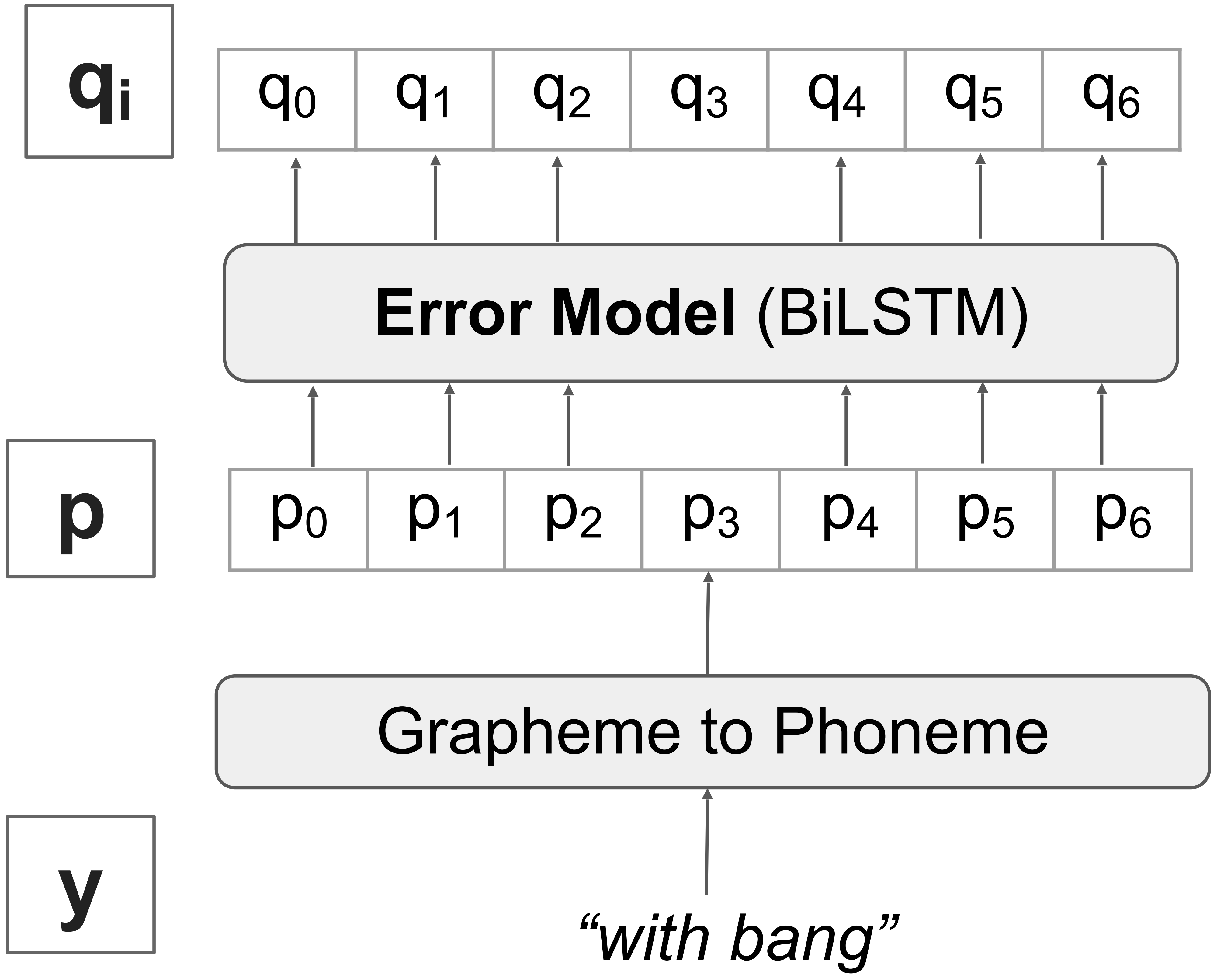
Prevents bias towards longer sentences

Scoring and Selecting Utterances using Error Model



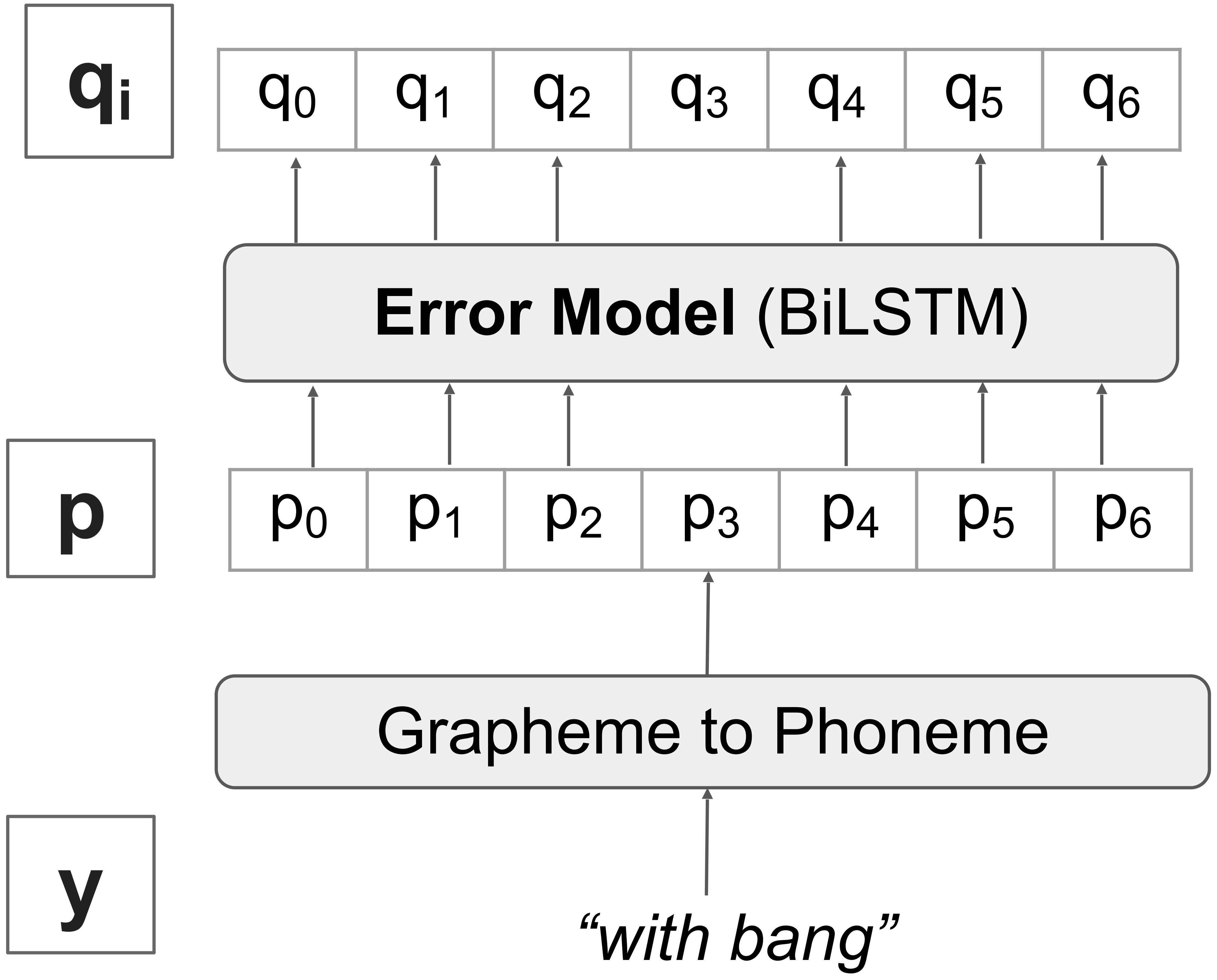
$$\mathrm{score}(\mathbf{y}) = \frac{1}{n}\sum_j q_j$$

Scoring and Selecting Utterances using Error Model



$$\text{score}(\mathbf{y}, \mathcal{Y}) = \frac{1}{n} \sum_{\pi \in \mathcal{P}} c_\pi(\mathcal{Y}, \mathbf{y}) \sum_{j:\mathrm{p}_j = \pi} q_j$$

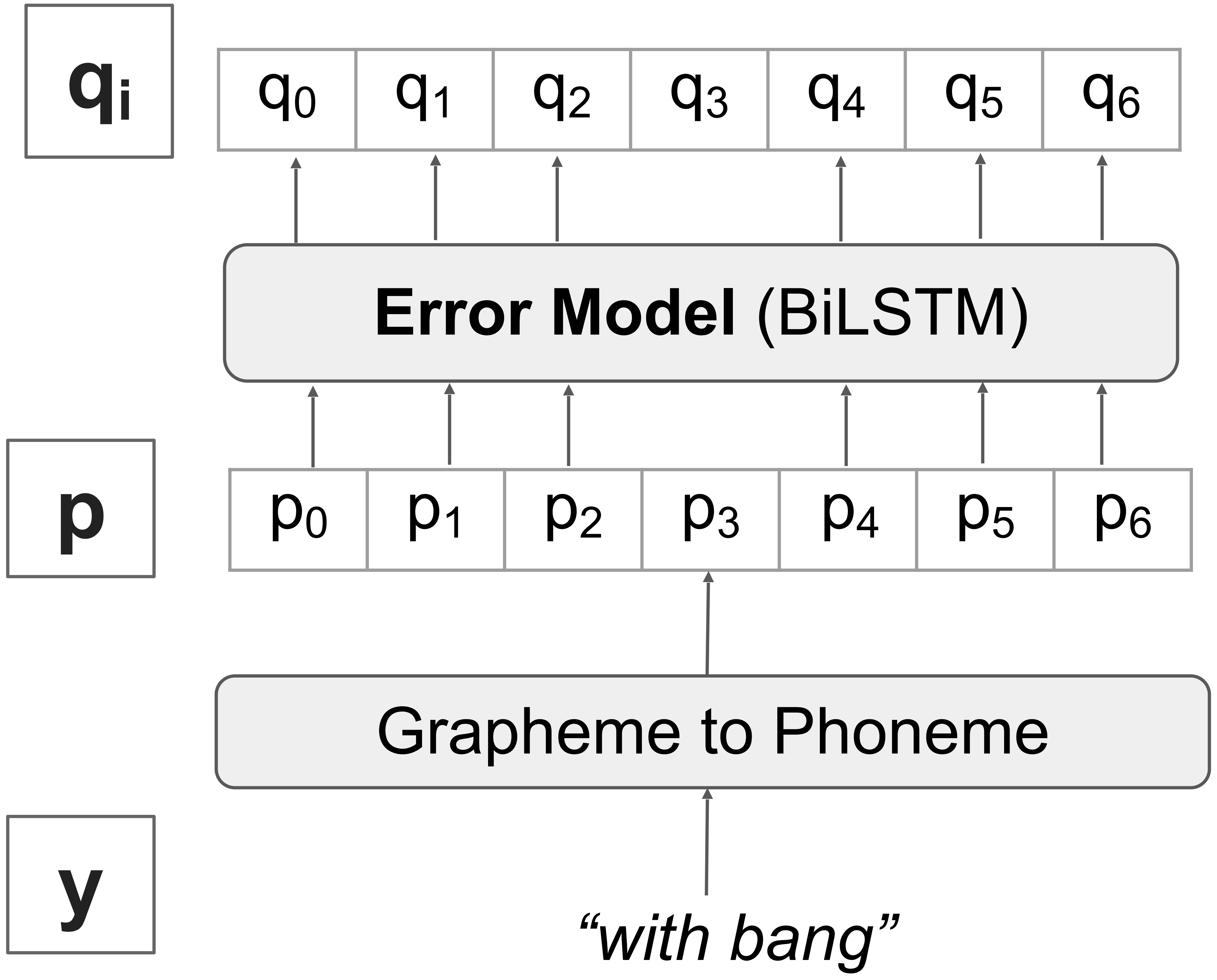
Scoring and Selecting Utterances using Error Model



$$\mathrm{score}(\mathbf{y}, \mathcal{Y}) = \frac{1}{n} \sum_{\pi \in \mathcal{P}} c_\pi(\mathcal{Y}, \mathbf{y}) \sum_{j:\mathrm{p}_j = \pi} q_j$$

$\mathcal{Y}$: Already selected sentences

Scoring and Selecting Utterances using Error Model

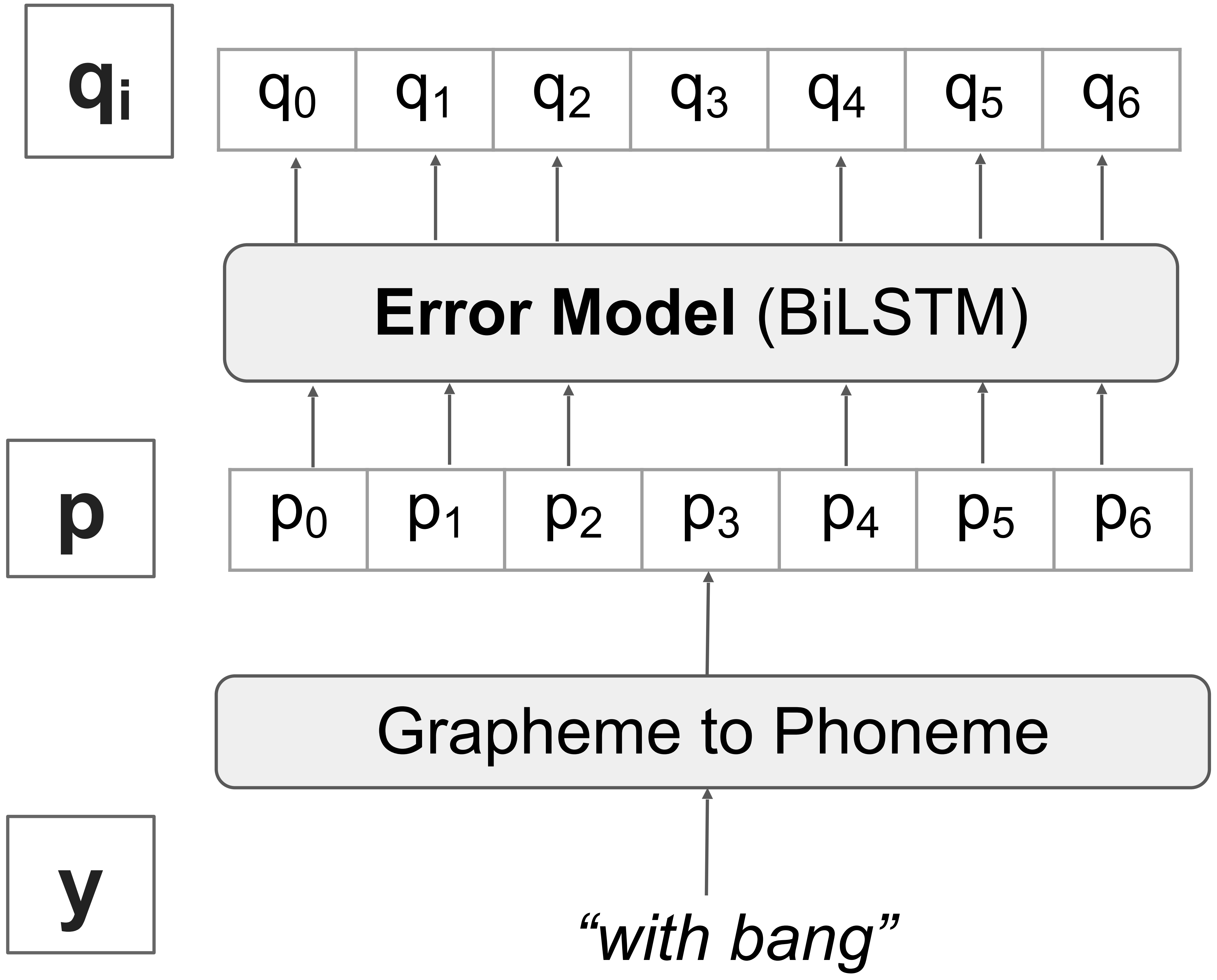


Diminishing returns for selecting a phone which already has a good count in the selected set

$$\mathrm{score}(\mathbf{y}, \mathcal{Y}) = \frac{1}{n} \sum_{\pi \in \mathcal{P}} c_{\pi}(\mathcal{Y}, \mathbf{y}) \sum_{j:\mathrm{p}_j = \pi} q_j$$

$\mathcal{Y}$: Already selected sentences

Scoring and Selecting Utterances using Error Model



Diminishing returns for selecting a phone which already has a good count in the selected set

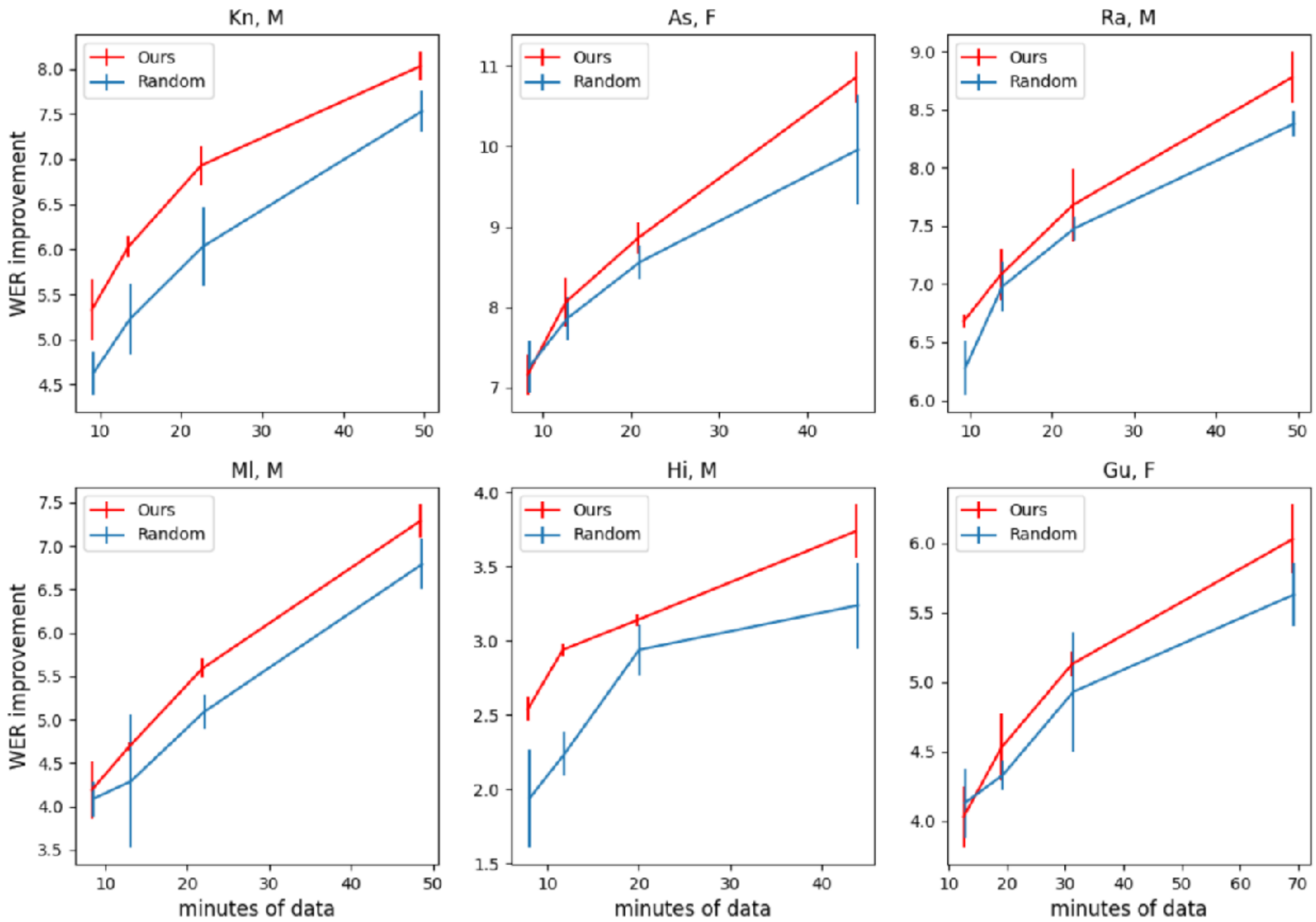
$$\text{score}(\mathbf{y}, \mathcal{Y}) = \frac{1}{n} \sum_{\pi \in \mathcal{P}} c_\pi(\mathcal{Y}, \mathbf{y}) \sum_{j : \mathrm{p}_j = \pi} q_j$$

$$\boxed{c_\pi(\mathcal{Y}, \mathbf{y}) = f(n_\pi(\mathcal{Y} \cup \mathbf{y})) - f(n_\pi(\mathcal{Y}))}$$

n_π : count of phoneme π

$$f(n_\pi) = 1 - \exp\left(-\frac{n_\pi}{\tau}\right)$$ (Submodular Function)

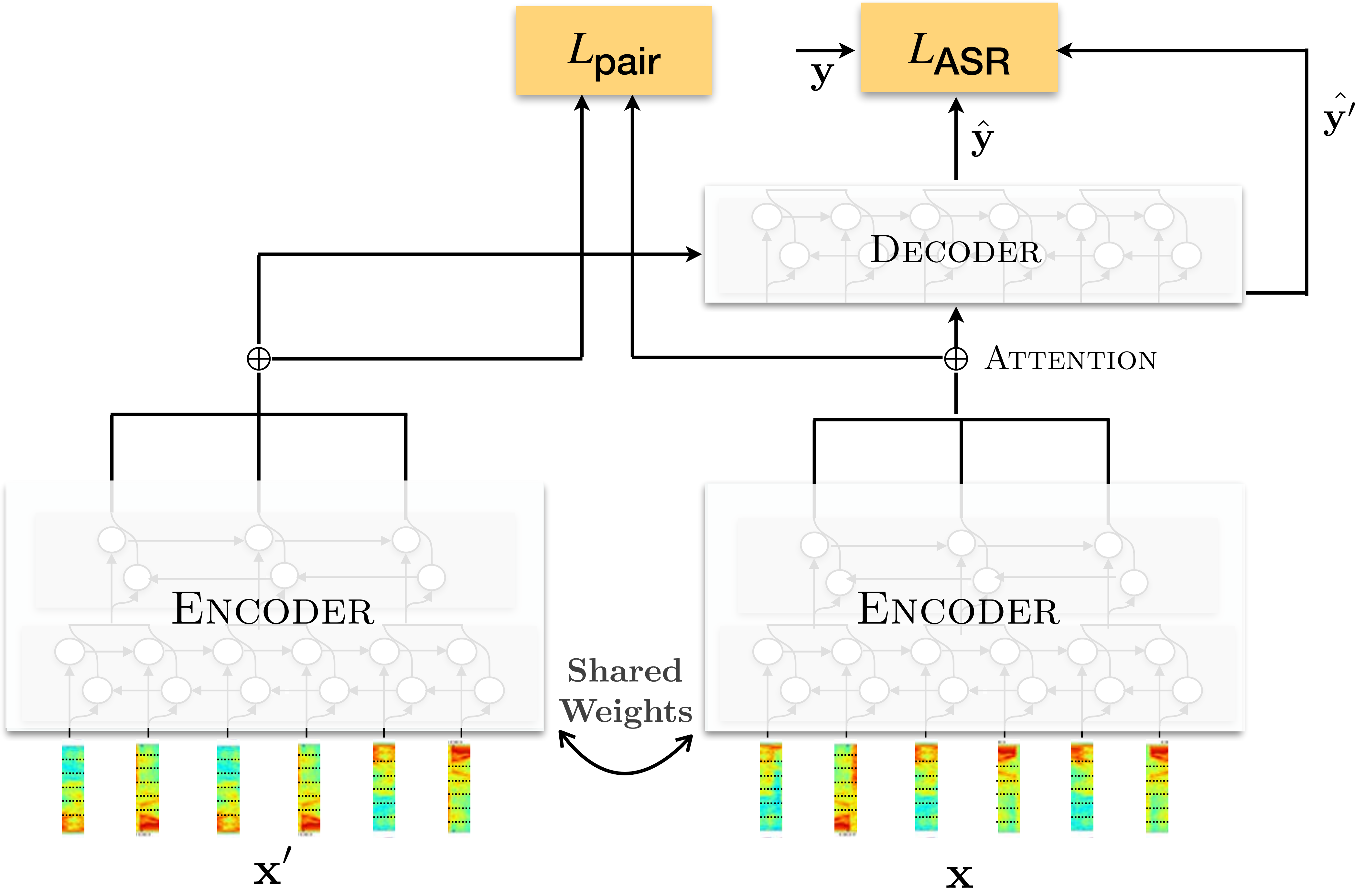
Fine-tuned ASR Using Selected Samples



Selection using our **error models** provide consistent gains over **random selection**

Accent-Agnostic Speech Recognition UJJ’20

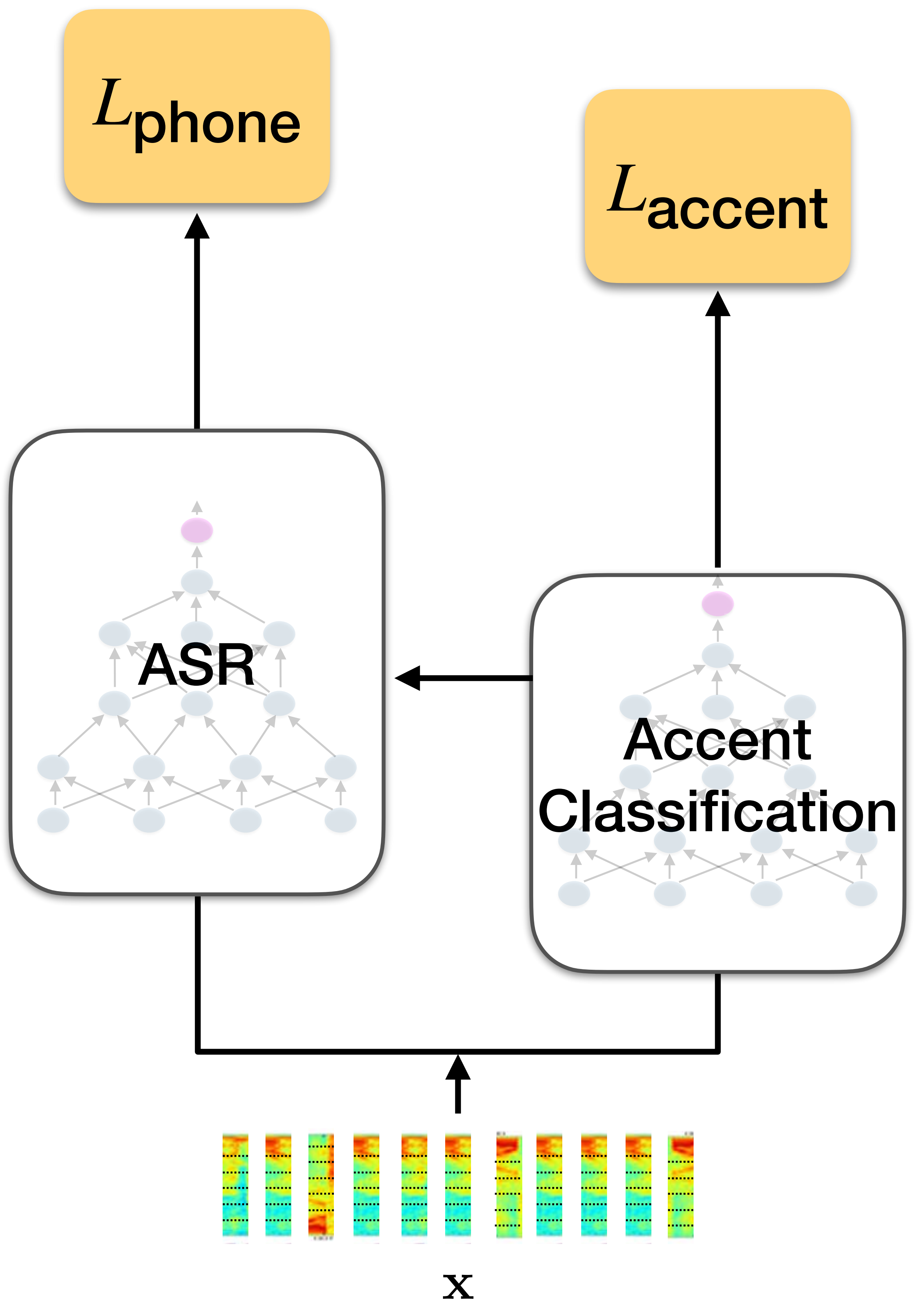
- Natural idea: Learn an internal representation that is accent-invariant
- *Coupled training* using parallel speech data (same text, differently accented speakers)
  - Leads to consistent performance improvements even on challenging Indian-accented samples
- But availability of parallel speech data is limited



UJJ’20 V. Unni, N. Joshi and P. Jyothi, “Coupled training of sequence-to-sequence models for accented speech recognition”, ICASSP 2020

Accent-Aware Speech Recognition JUJ'18

- Alternate approach: Learn to handle different accents differently
- Plan: Actively extract and use "accent information"
  - Accent information obtained in two ways:
    - Accent embedding produced by an accent classifier trained separately
    - Tapped from an accent classifier trained alongside ASR (multi-task training)
  - And fed into an appropriate layer in the ASR network
- Significantly lower error rates compared to a multi-accent baseline:
  - 15% on seen accents
  - 9% on a new accent



JUJ'18 A. Jain, M. Upreti and P. Jyothi, "Improved Accented Speech Recognition using Accent Embeddings and Multi-task Learning", Interspeech 2018

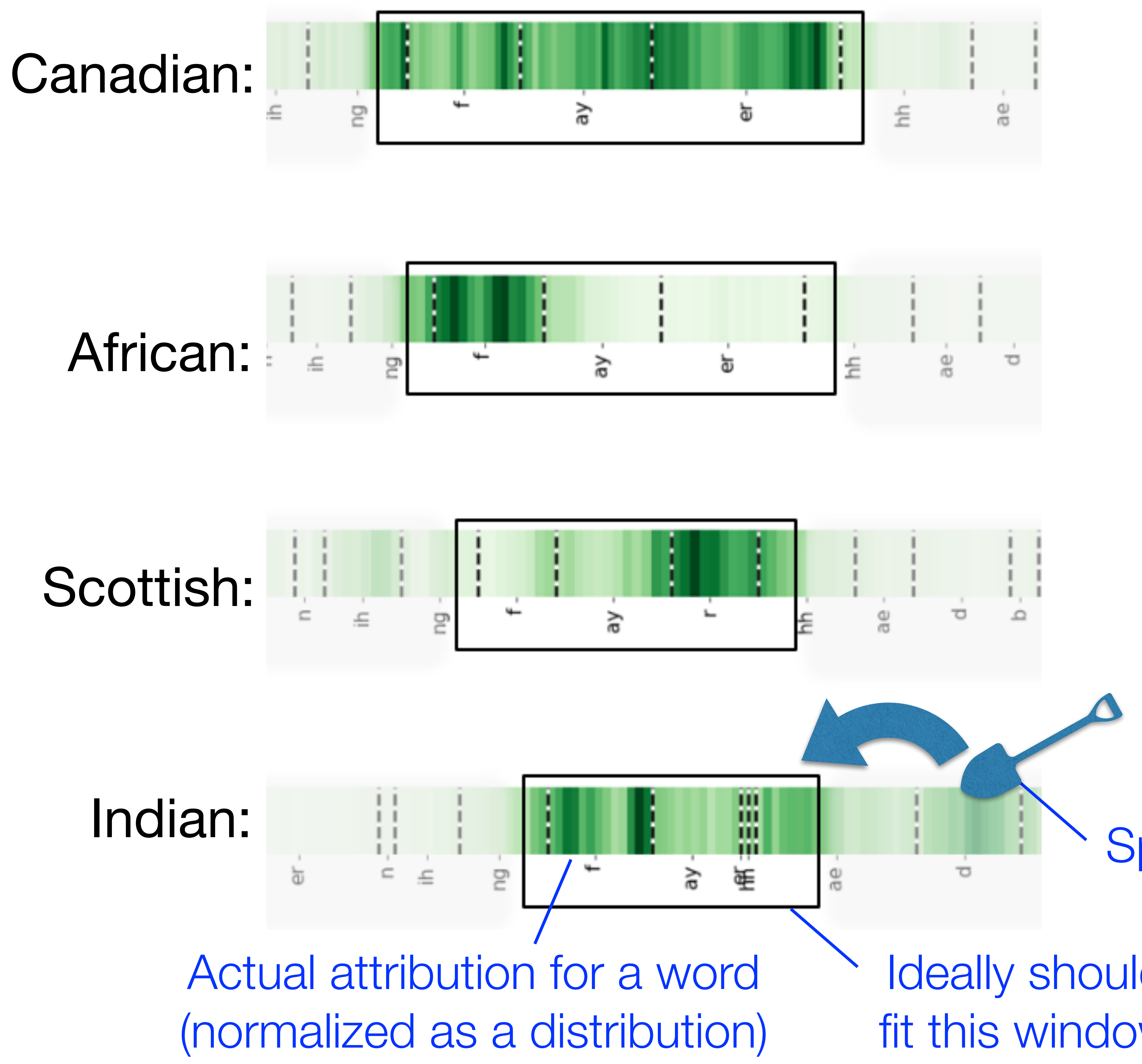
Understanding Accent in Neural Networks PJ'20

- How do neural networks handle accents?
- A study of DeepSpeech2 using different measures and tools
  - **Gradients based:** While outputting each word, how well the network "focuses" on the correct segment.
  - **Information in layers:** Amount of information that representations at various layers carry about the accent, and for each accent, about the phones.
    - Information theoretic: Measured using mutual information (after clustering the representations).
    - Classifier based: Measured using the accuracy of a classifier that takes the representations as inputs.
- Improving ASR systems using such analysis while designing them

PJ'20 A. Prasad, P. Jyothi, "How Accents Confound: Probing for Accent Information in End-to-End Speech Recognition Systems", ACL 2020

Understanding Accent Information in Neural Networks

`



Accent	EMD
Canadian	40.9
US	42.6
African	44.3
English	44.3
Scottish	43.3
Australian	45.9
Indian	50.3

Lowest

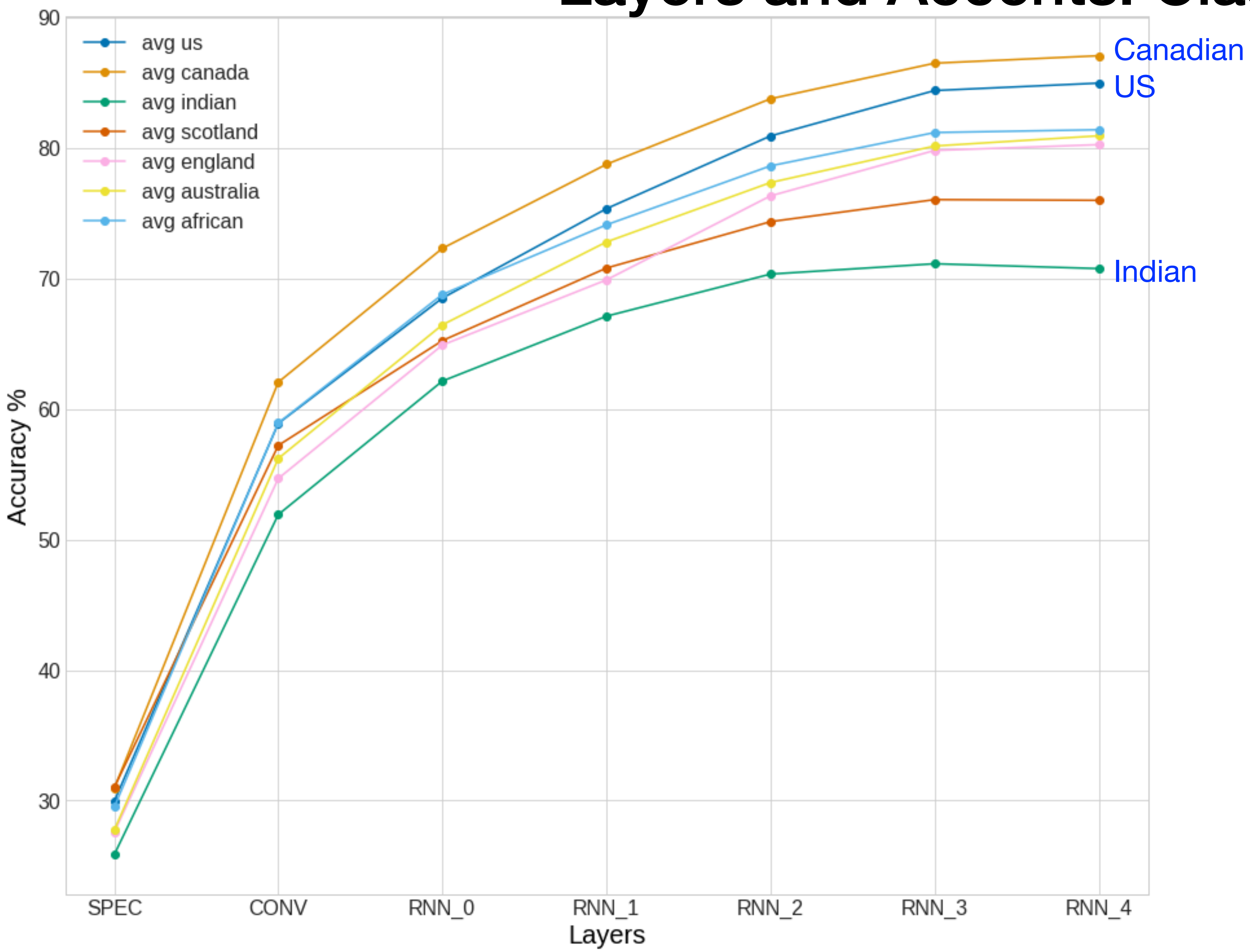
Highest

Spillage/gap measured using
Earth Mover Distance

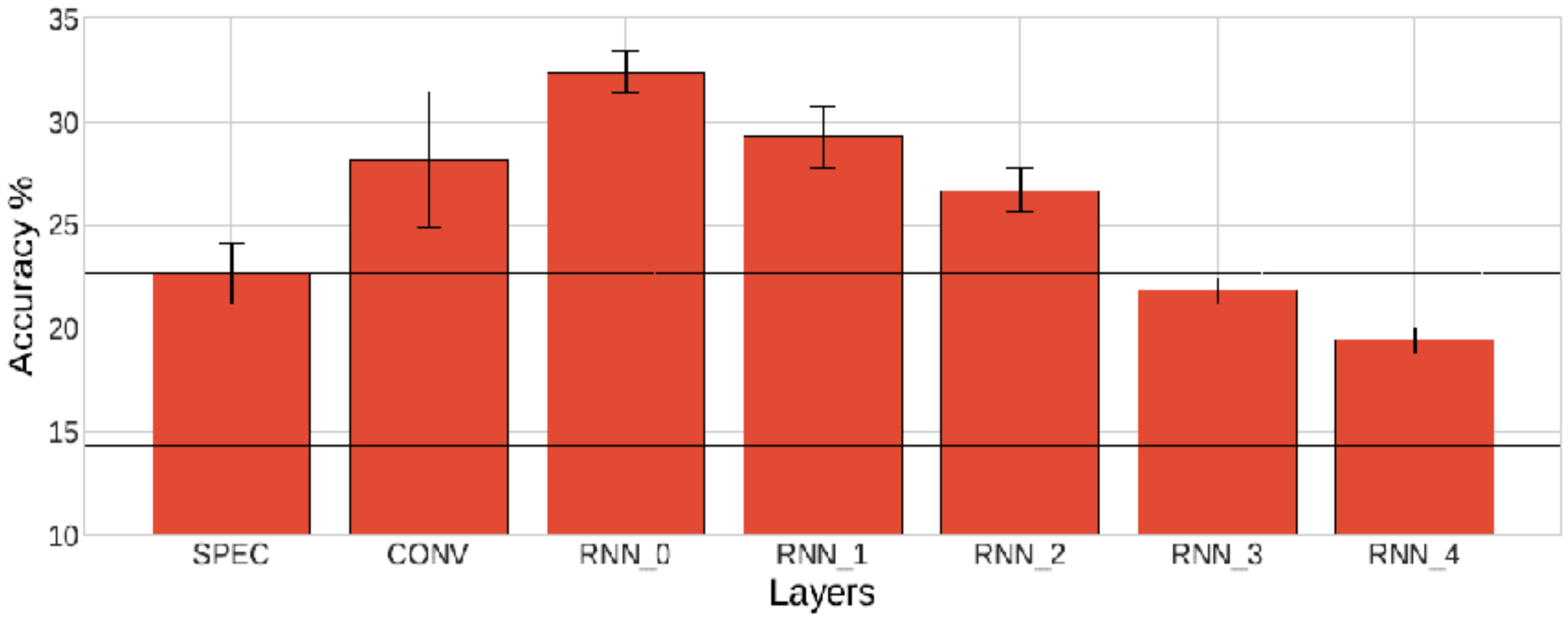
A. Prasad, P. Jyothi, “How Accents Confound: Probing for Accent Information in End-to-End Speech Recognition Systems”, ACL 2020

Understanding Accent Information in Neural Networks PJ'20

Layers and Accents: Classifier-Based Analysis



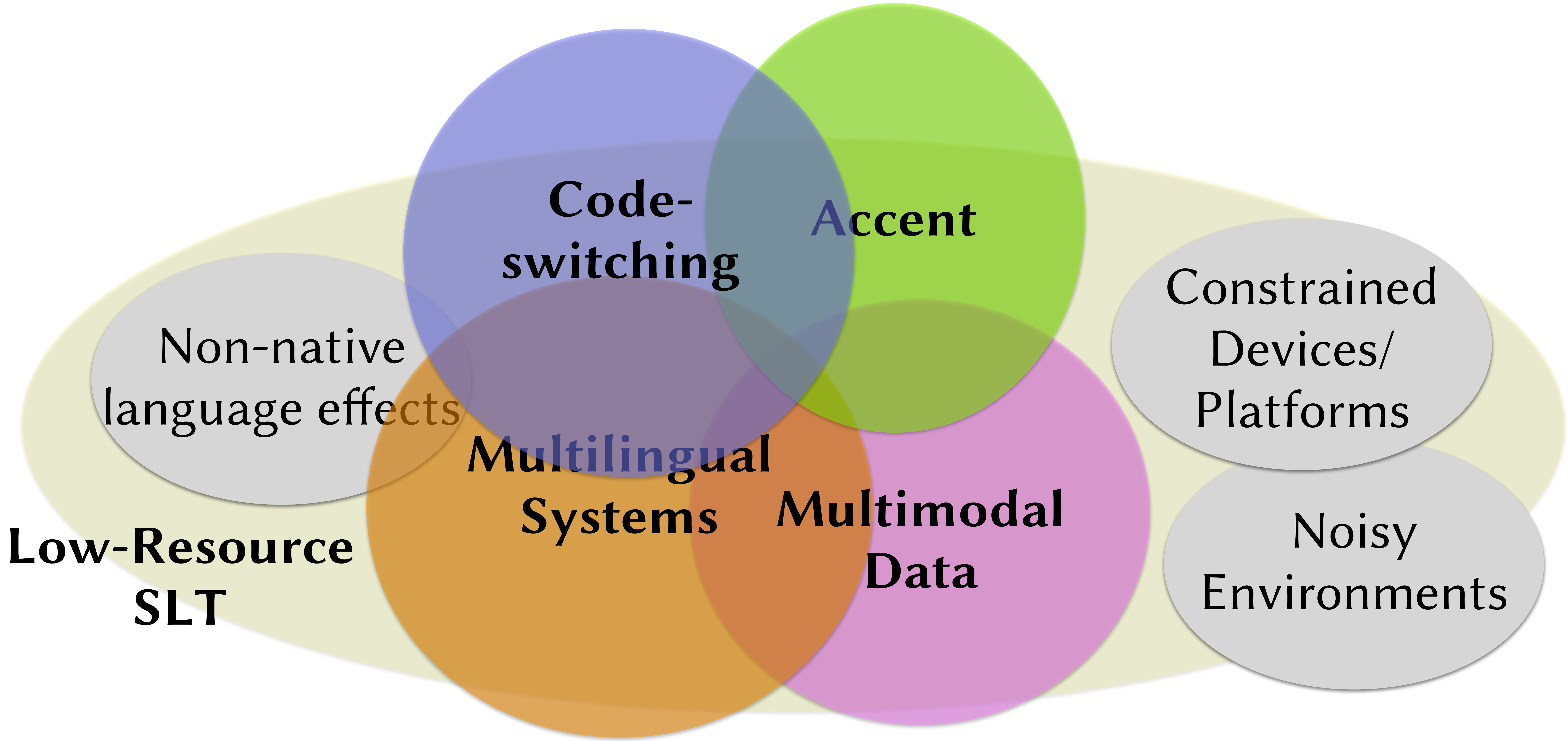
Accuracy of phone probes across layers



Accuracy of accent probes across layers

PJ'20 A. Prasad, P. Jyothi, "How Accents Confound: Probing for Accent Information in End-to-End Speech Recognition Systems", ACL 2020

Speech and Language Technologies (SLT) for India



Code-Switching

- Switching between different languages within/across sentences

  Piya Tose Naina Laage का Amazing Rendition Deliver किया इस Audition पे

  - Widely prevalent in multilingual countries like India
- An emerging sub-area in SLT
- Just treat it like a new language?
  - Hard to get access to large amounts of code-switched data
  - Large diversity in how code-switching manifests

  But laughter therapy ने मेरी life बदल दी really

  But laughter therapy ने really में मेरी life change कर दी

  पर हंसी therapy ने मेरी life बदल दिया वास्तव में

Dual Language Model

- Recall code-switching issues:

  Hard to get access to large amounts of code-switched data

  Large diversity in how code-switching manifests

- Two high-level ideas for fixing them:

  Should exploit monolingual data in each language

  Should model both languages separately in addition to modeling how switching occurs

  synergistic

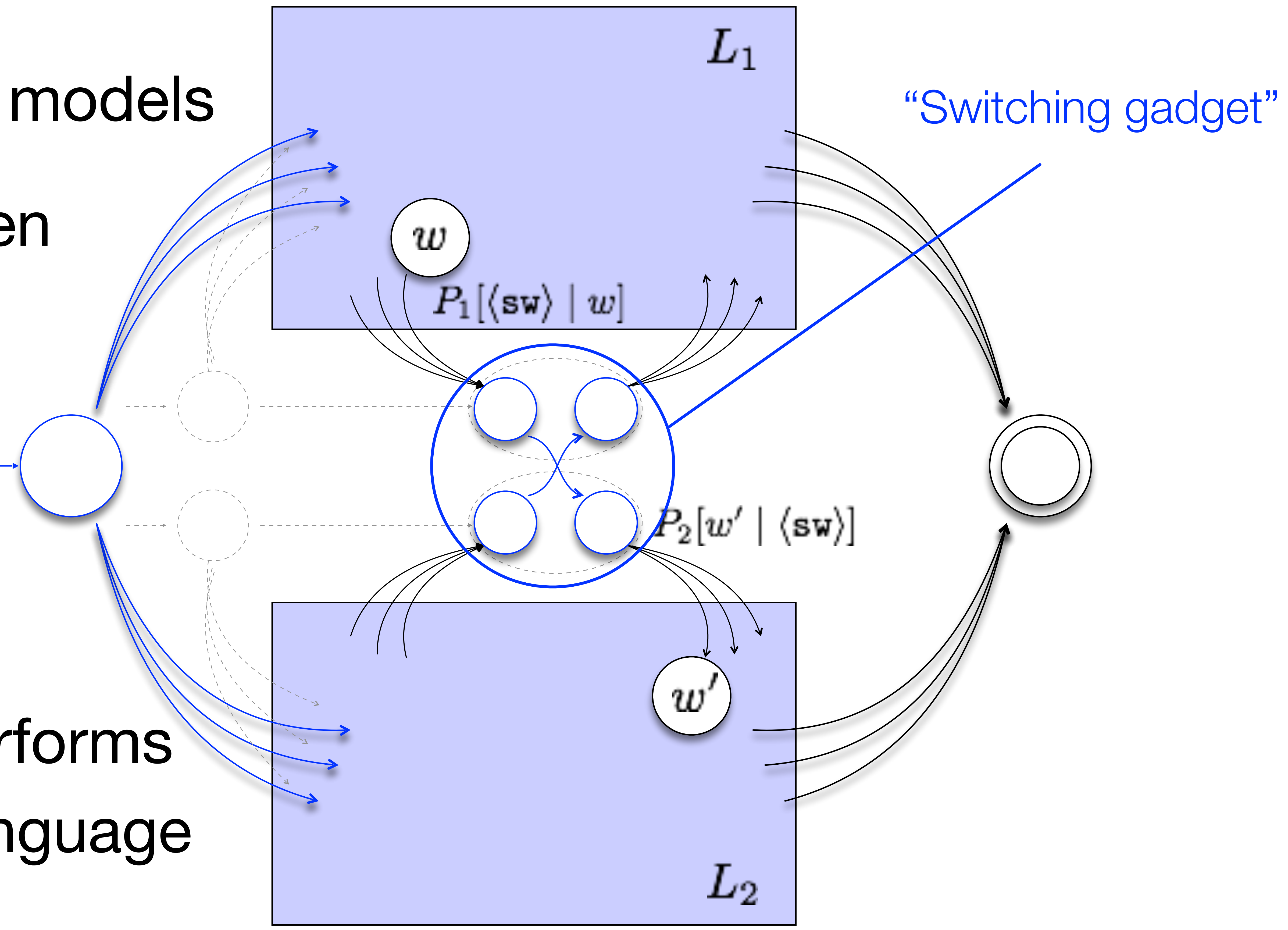
- Dual Language Models
  - n-gram language model GPJ'18a
  - Recurrent Neural Network model GPJ'18b

GPJ'18a Garg et al., "Dual Language Models for Code Switched Speech Recognition", INTERSPEECH 2018

GPJ'18b Garg et al., "Code-switched Language Models Using Dual RNNs and Same-Source Pretraining", EMNLP 2018

Dual Language Model : With n-grams GPJ'18a

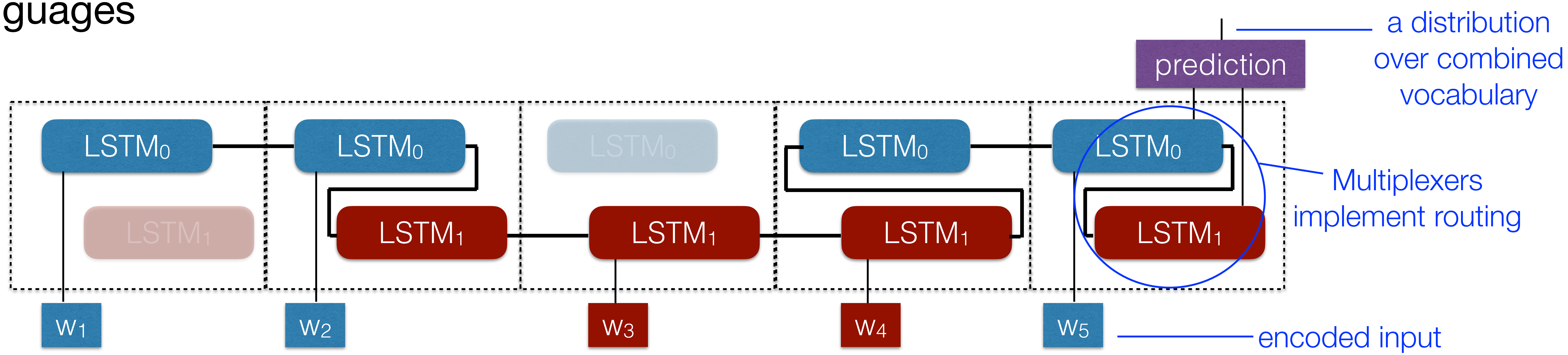
- *n*-gram language models, represented as Weighted Finite-State Transducers (WFST)
  - Standard for "conventional" ASR models
  - Can also be integrated into neural network models
- We combine two such LMs, switching between them via a "switching gadget"
  - Switches with state-specific probabilities, which can be learnt from a relatively small amount of data
- Even without using mono-lingual text, out-performs a monolithic LM that treats code-switched language as a "new" language



GPJ'18a Garg et al., "Dual Language Models for Code Switched Speech Recognition", INTERSPEECH 2018

Dual Language Model : With RNNs GPJ'18b

- Could we do better?
  - Neural network models tend to out-perform *n*-gram models
  - Also, the *n*-gram Dual LM dropped all contextual information during a switch
- An RNN that has two different units (LSTM cells) for handling sequences in the two different languages

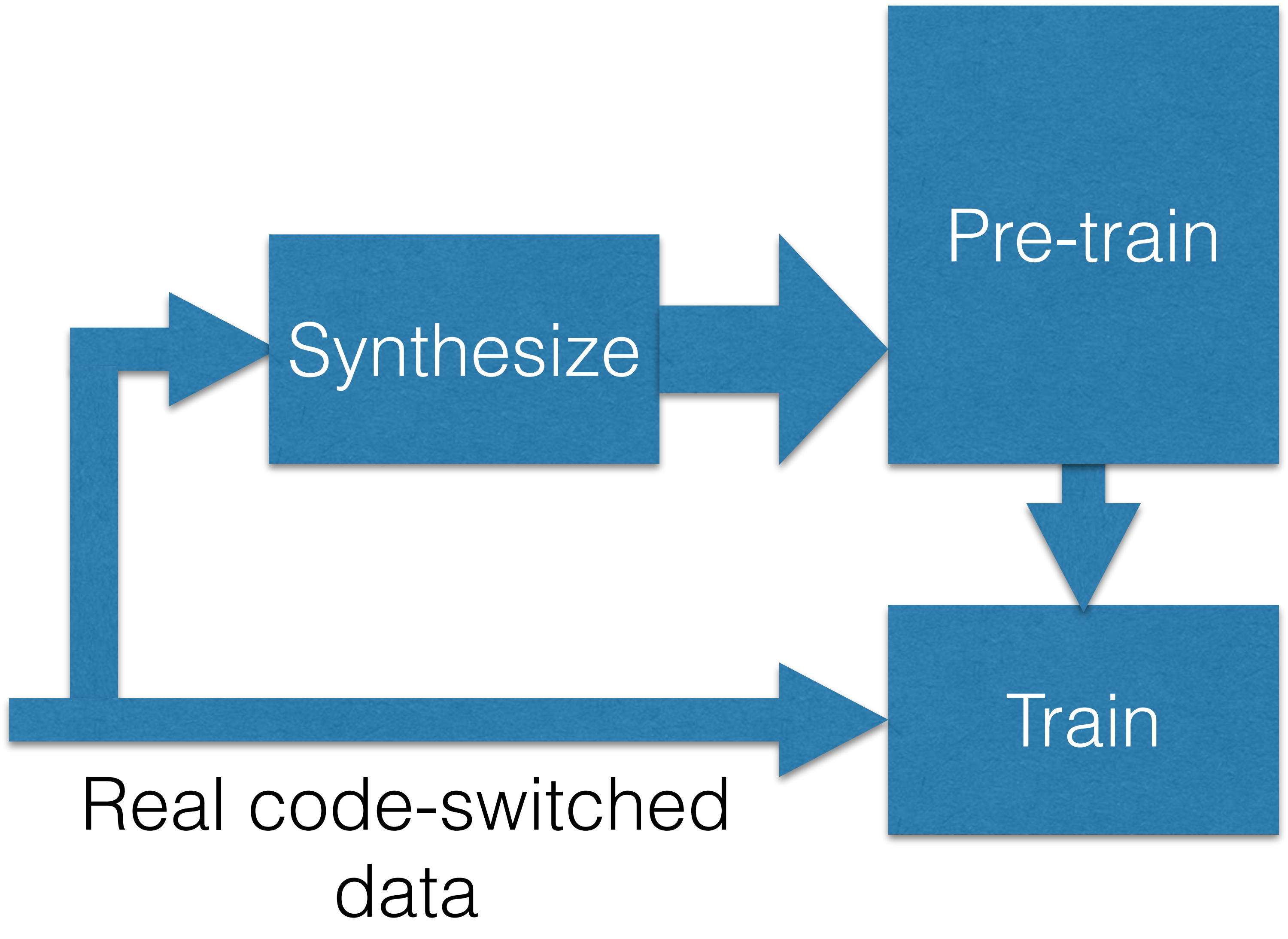


- Can train using mono-lingual text and code-switched text
- A problem: We don't have enough code-switched text

GPJ'18b Garg et al., "Code-switched Language Models Using Dual RNNs and Same-Source Pretraining", EMNLP 2018

Dual Language Model : With RNNs GPJ'18b And Same-Source Pre-Training

- A problem: We don't have enough code-switched text
- Solution: Use "synthetic data" (possibly of lower quality) to *pre-train* the RNN
- But how do we synthesize code-switched data?
  - Use a generator trained on the (low amounts of) real data
  - Note: Same source used for both training and for generating data for pre-training
    - Works!

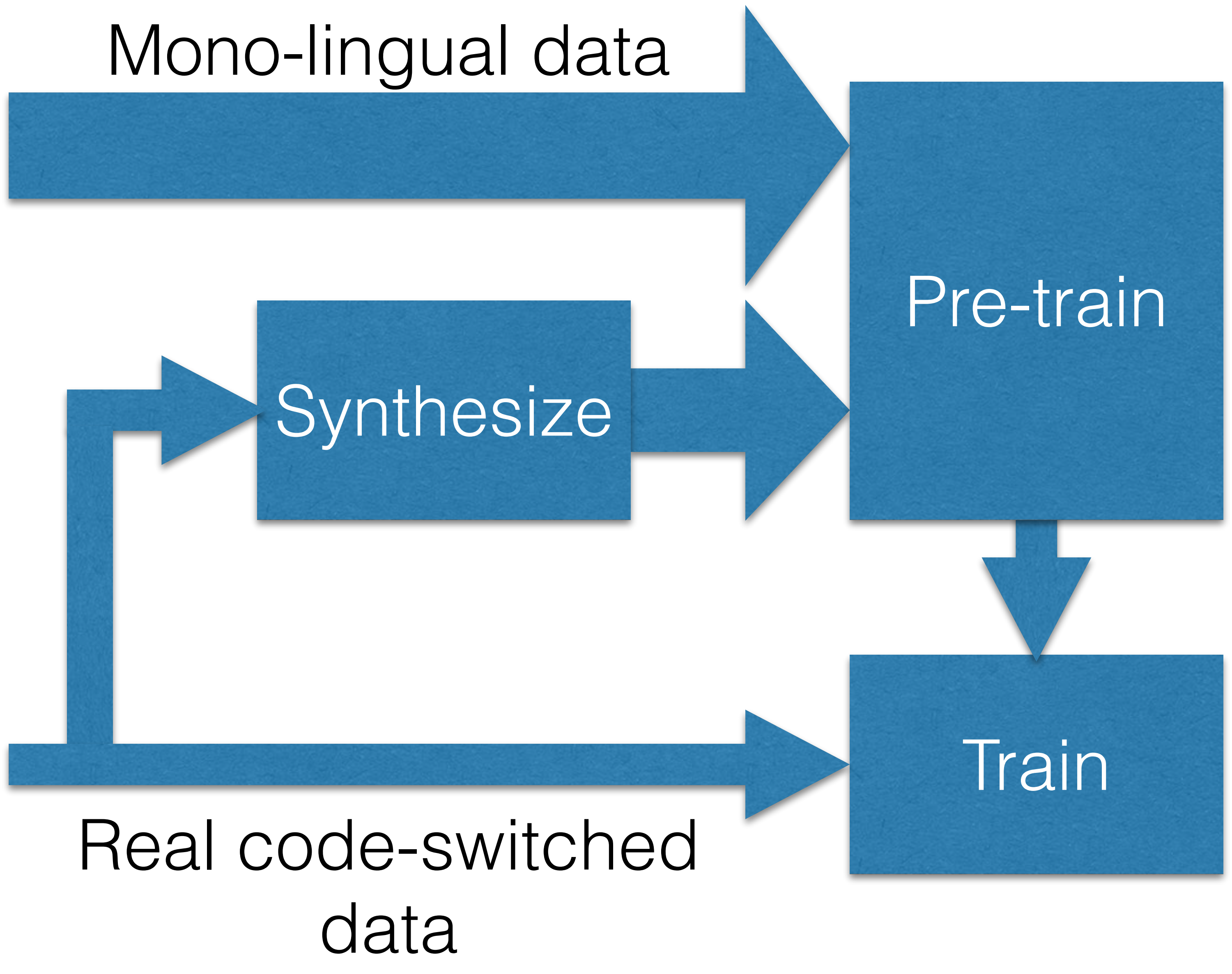


perplexity (low is good)	RNN	+Dual
RNN	68.2	66.3
+Synth	63.8	**63.6**

GPJ'18b Garg et al., "Code-switched Language Models Using Dual RNNs and Same-Source Pretraining", EMNLP 2018

Dual Language Model : With RNNs GPJ'18b And Same-Source Pre-Training

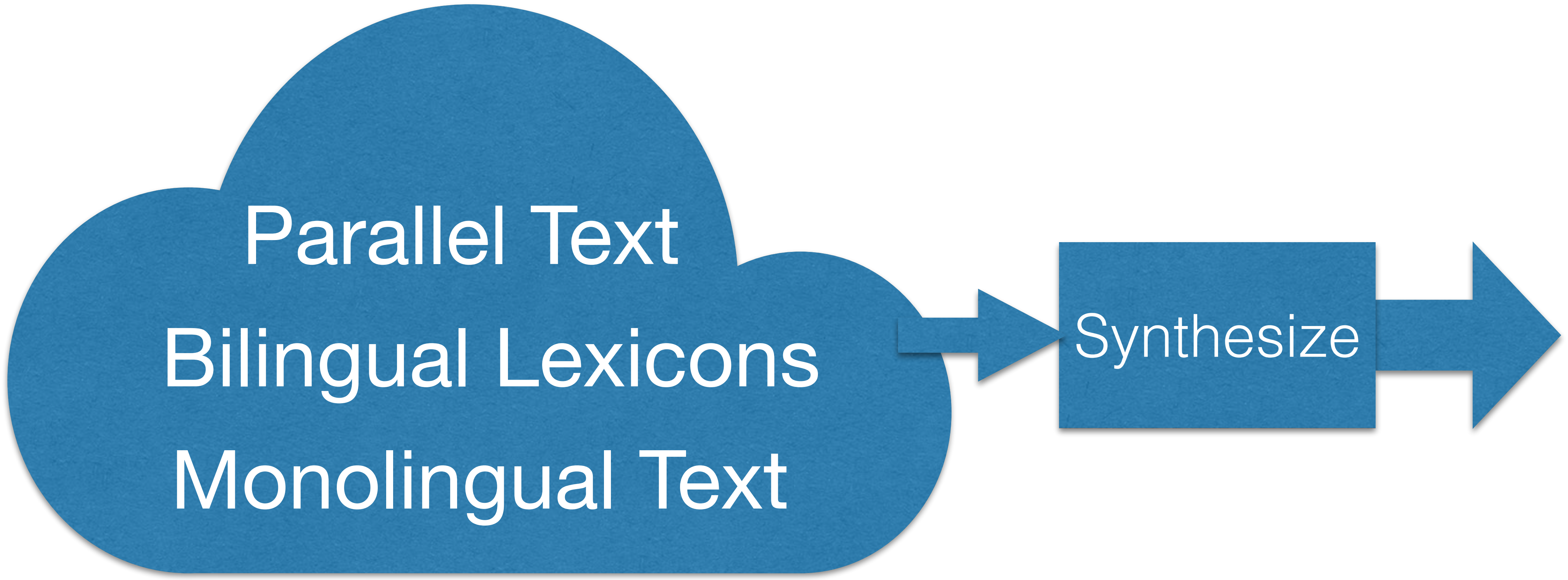
- A problem: We don't have enough code-switched text
- Solution: Use "synthetic data" (possibly of lower quality) to *pre-train* the RNN
- But how do we synthesize code-switched data?
  - Use a generator trained on the (low amounts of) real data
  - Note: Same source used for both training and for generating data for pre-training
    - Works!
- Can effectively exploit mono-lingual data too



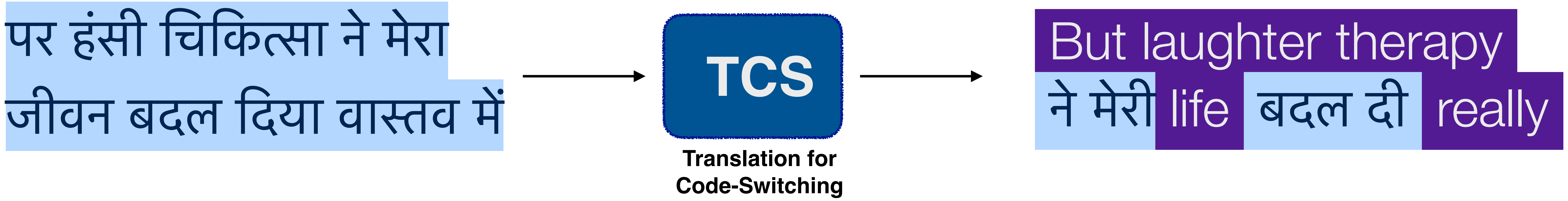
perplexity (low is good)	RNN	+Dual
RNN	59.0	59.0
+Synth	55.7	**55.6**

Generating Code-switched Text

- Generating synthetic, but realistic code-switched text is an important problem on its own



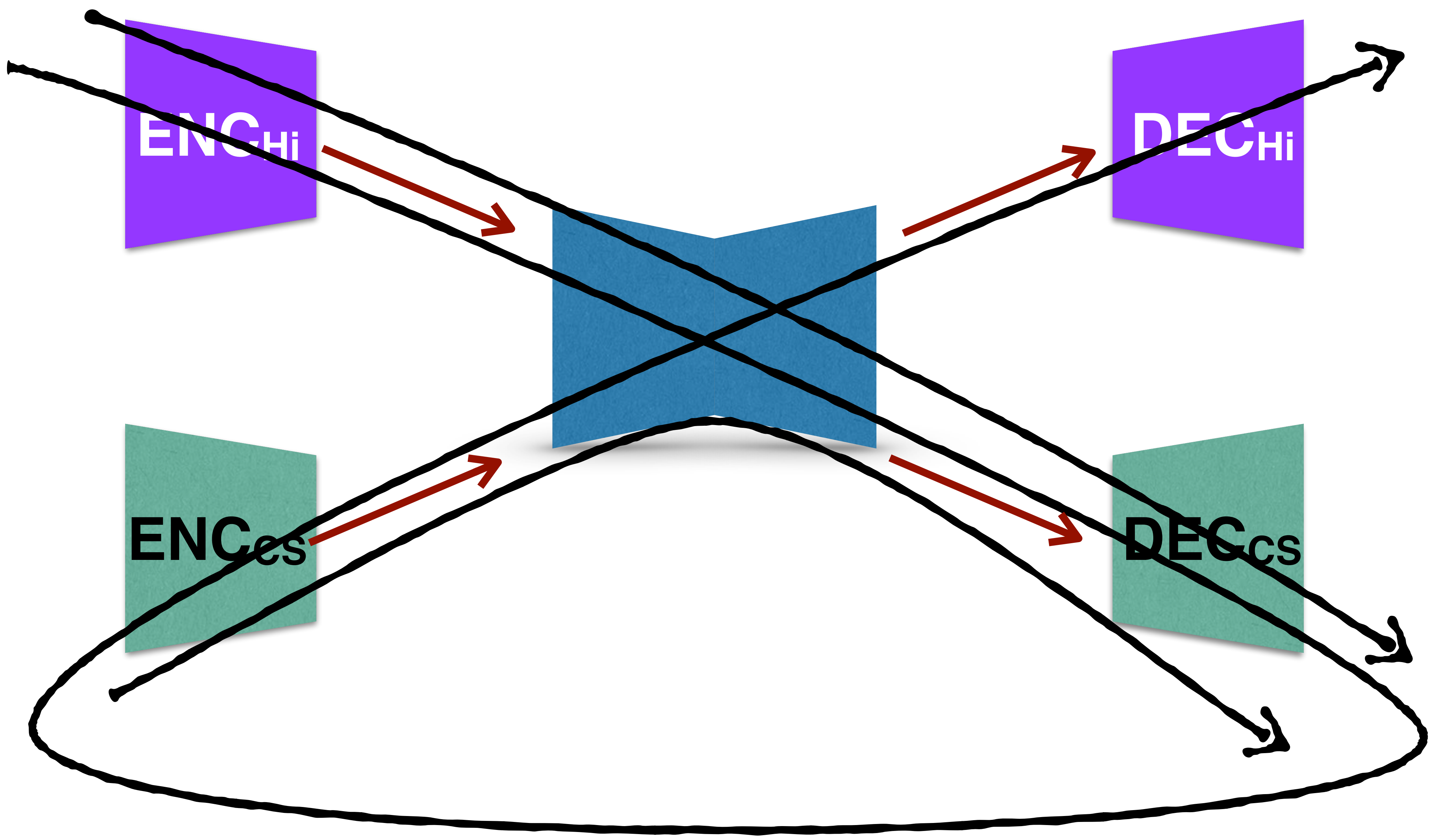
  - Can we leverage more resources?
  - A different idea: Treat it as a translation task! TKJ'21
    - E.g., Convert a monolingual Hindi sentence to a Hindi-English sentence



TKJ'21 Tarunesh et al., "From Machine Translation to Code-Switching: Generating High-Quality Code-Switched Text", ACL 2021

Generating Code-switched Text: Translation to Code-Switching TKJ'21

- Based on an unsupervised MT architecture LXX'18
  - Can use monolingual text and code-switched text. Parallel text is optional for training.
  - We also employ (simplistically generated) synthetic code-switched text
    - LEX: Use a bilingual lexicon to replace a random Hindi word by its English translation
    - EMT: Replace embedded sentence clauses or subordinate clauses in English sentences with Hindi translations
  - Supervised version TCS(S) using a new dataset for parallel code-switched & Hindi text

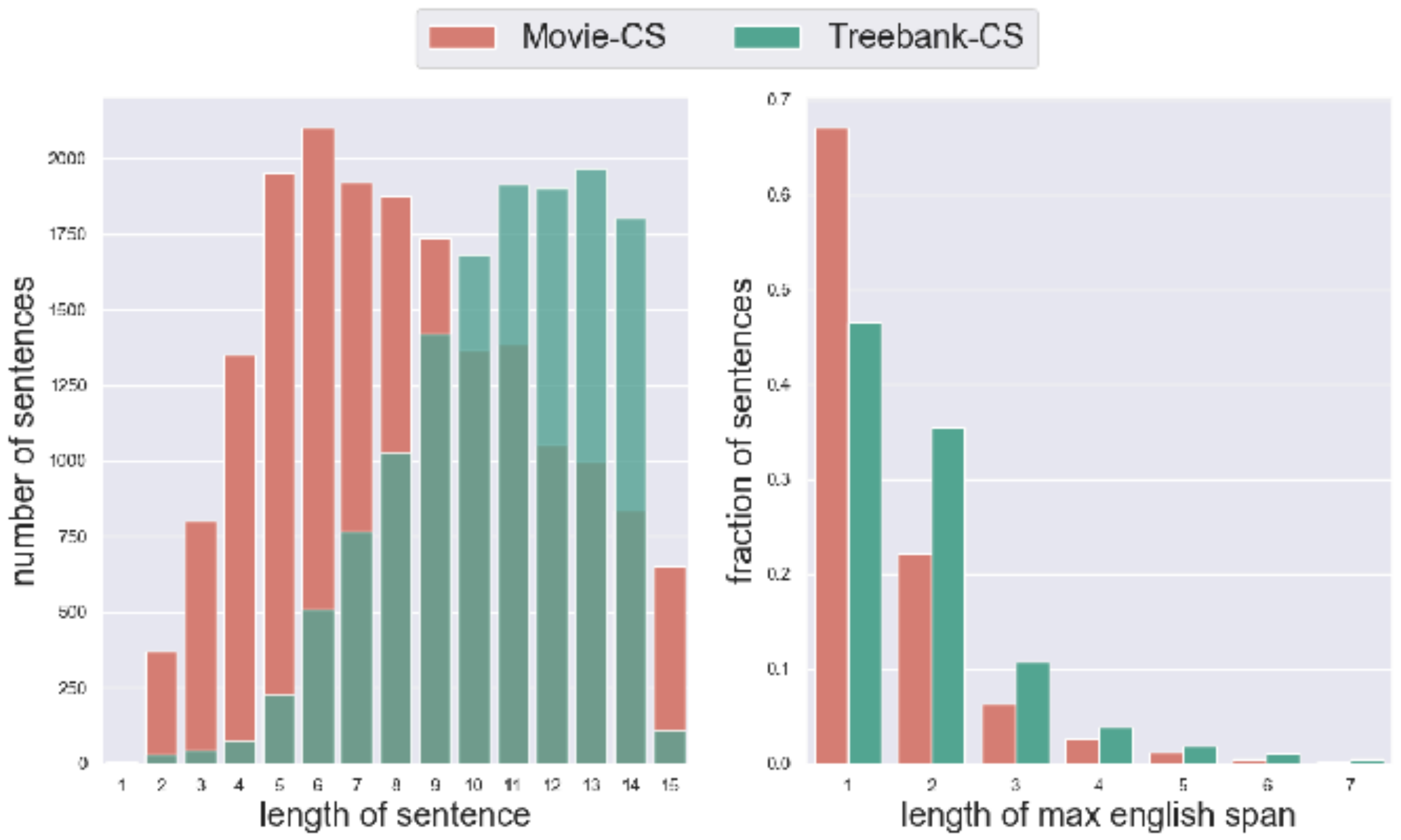


TKJ'21 Tarunesh et al., "From Machine Translation to Code-Switching: Generating High-Quality Code-Switched Text", ACL 2021

LXX'18 "Unsupervised Machine Translation Using Monolingual Corpora Only", Lample, G. et al., ICLR 2018

Generating Code-switched Text: Translation to Code-Switching

A new Hindi-English CS Dataset



- Contains 21K+2.5K train+test instances
- Partitioned into two subsets:
  Movie-CS and *Treebank-CS*
- Many of the CS sentences are crowdsourced using MTurk
- For sentences in *Treebank-CS*, Turkers were asked to translate at least one Hindi chunk into English

Human Evaluation

Human Evaluation

Method	Syntactic	Semantic	Naturalness
Real	4.47±0.73	4.47±0.76	4.27±1.06
TCS (S)	4.21±0.92	4.14±0.99	3.77±1.33
TCS (U)	4.06±1.06	4.01±1.12	3.58±1.46
EMT	3.57±1.09	3.48±1.14	2.80±1.44
LEX	2.91±1.11	2.87±1.19	1.89±1.14

Syntactic Correctness: Is the sentence grammatically valid?

Semantic Correctness: Is the sentence semantically meaningful?

Naturalness: Does the sentence look naturally code-switched?

TCS: Example Sentences

नहीं मैं तुमसे बहुत प्यार करता हूँ सच में लेकिन सिर्फ एक दोस्त की तरह → TCS → नहीं i love you very much सच में but सिर्फ एक friend की तरह

क्या बात है तुमने आखरी बार कब पार्टी की थी → TCS → क्या बात है तुमने last time party कब की थी

स्कूलों में तो नियमित रूप से सुरक्षा अभ्यास कराए जाने लगे हैं → TCS → schools में तो regularly security practice किये जाने लगे हैं

Evaluation using Objective Measures TKJ'21

Objective Measures

Measure	LEX	EMT	TCS(S)
BLEU	15.23	17.73	**43.15**
LM Perplexity	332.66	276.56	**254.37**
GLUECoS - NLI	58.67	58.96	**59.57**
GLUECoS - SA	58.40	58.79	**59.39**
BERT-Score	0.785	0.633	**0.813**
BERT-Classifier	96.52	97.83	**88.62**

TKJ'21 Tarunesh et al., "From Machine Translation to Code-Switching: Generating High-Quality Code-Switched Text", ACL 2021

Recall Diversity in Code-switching

- Diversity in code-switching caused by:
  - Sociolinguistic factors. E.g., 1st generation immigrants vs. younger immigrants
  - Formality in the rendered text. E.g., news vs. social media posts

But laughter therapy ने मेरी life बदल दी

But laughter therapy ने really में मेरी life change कर दी

पर हंसी therapy ने मेरी life बदल दिया वास्तव में

- We focus on three dimensions of diversity in code-switched text:

Code-mixing Index (CMI): Ratio of L1/L2 words

0.29	Gracias for the lovely gift, está awesome!
0.14	Gracias por el hermoso regalo, está awesome!

Switch-point Index (SPI): Freq. of L1/L2 switches

0.50	Gracias for the lovely gift, está awesome!
0.33	Thanks por el hermoso regalo, it's awesome!

Formality: Style, tone, choice of words

formal	इस पर comprehensive plan prepare की जा रही है
informal	इस पे detailed planning ready की जा रही है

Recall Diversity in Code-switching

- Diversity in code-switching caused by:
  - Sociolinguistic factors. E.g., 1st generation immigrants vs. younger immigrants
  - Formality in the rendered text. E.g., news vs. social media posts

But laughter therapy ने मेरी life बदल दी

But laughter therapy ने really में मेरी life change कर दी

पर हंसी therapy ने मेरी life बदल दिया वास्तव में

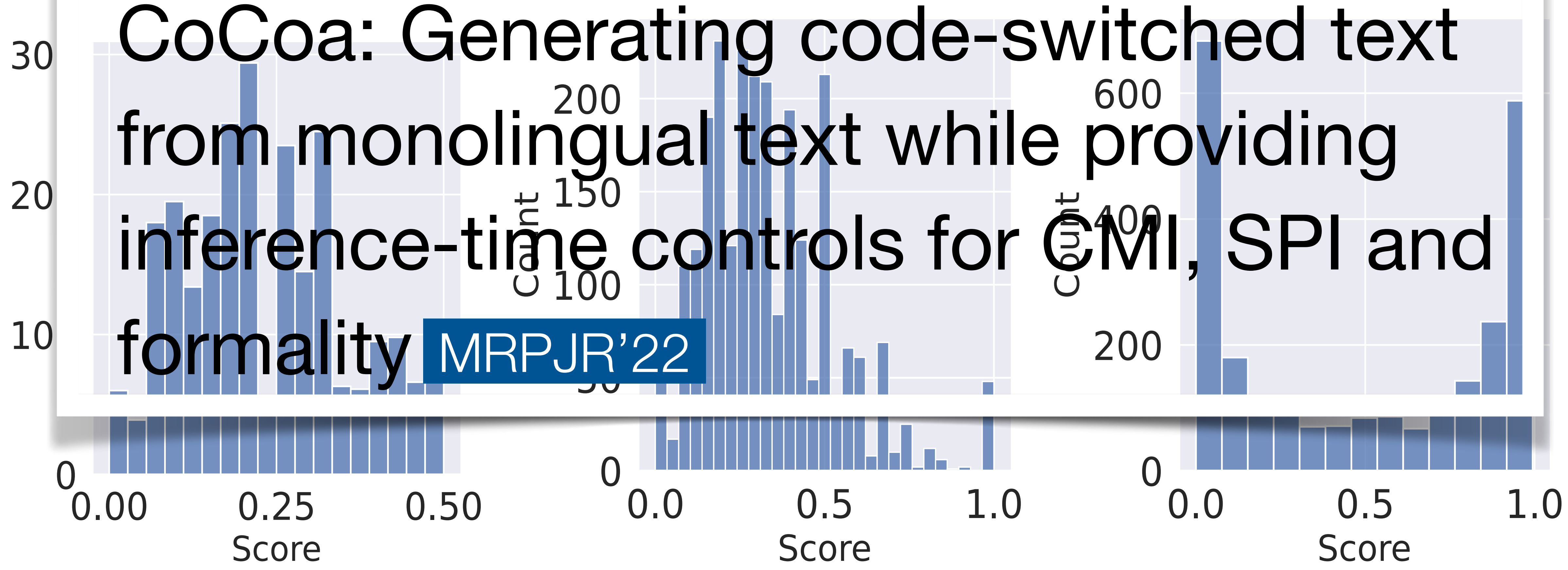
- We focus on three dimensions of diversity in code-switched text:

Code-mixing Index (CMI): Ratio of L1/L2 words

Switch-point Index (SPI): Freq. of L1/L2 switches

Formality: Style, tone, choice of words

CoCoa: Generating code-switched text from monolingual text while providing inference-time controls for CMI, SPI and formality MRPJR'22



(a) CMI (b) SPI (c) Formality

MRPJR'22 Mondal et al., "CoCoa: An Encoder-Decoder Model for Controllable Code-switched Generation", EMNLP 2022

Why is Diversity Computationally Important?

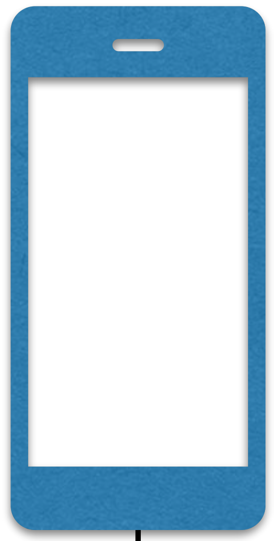
Understanding Diversity

Set an alarm for 8 am tomorrow

Kal subah 8 baje ka alarm set karo

Please kal 8 am ka alarm laga dein

Tomorrow 8 am alarm set kar dijiye



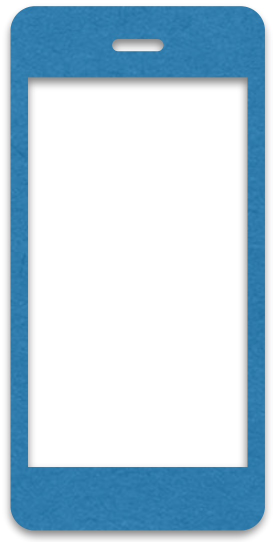
create_alarm (datetime (next_day 8 AM))

Generating Diversity

Set an alarm for 8 am tomorrow

Please kal 8 am ka alarm laga dein

Tomorrow 8 am alarm set kar dijiye



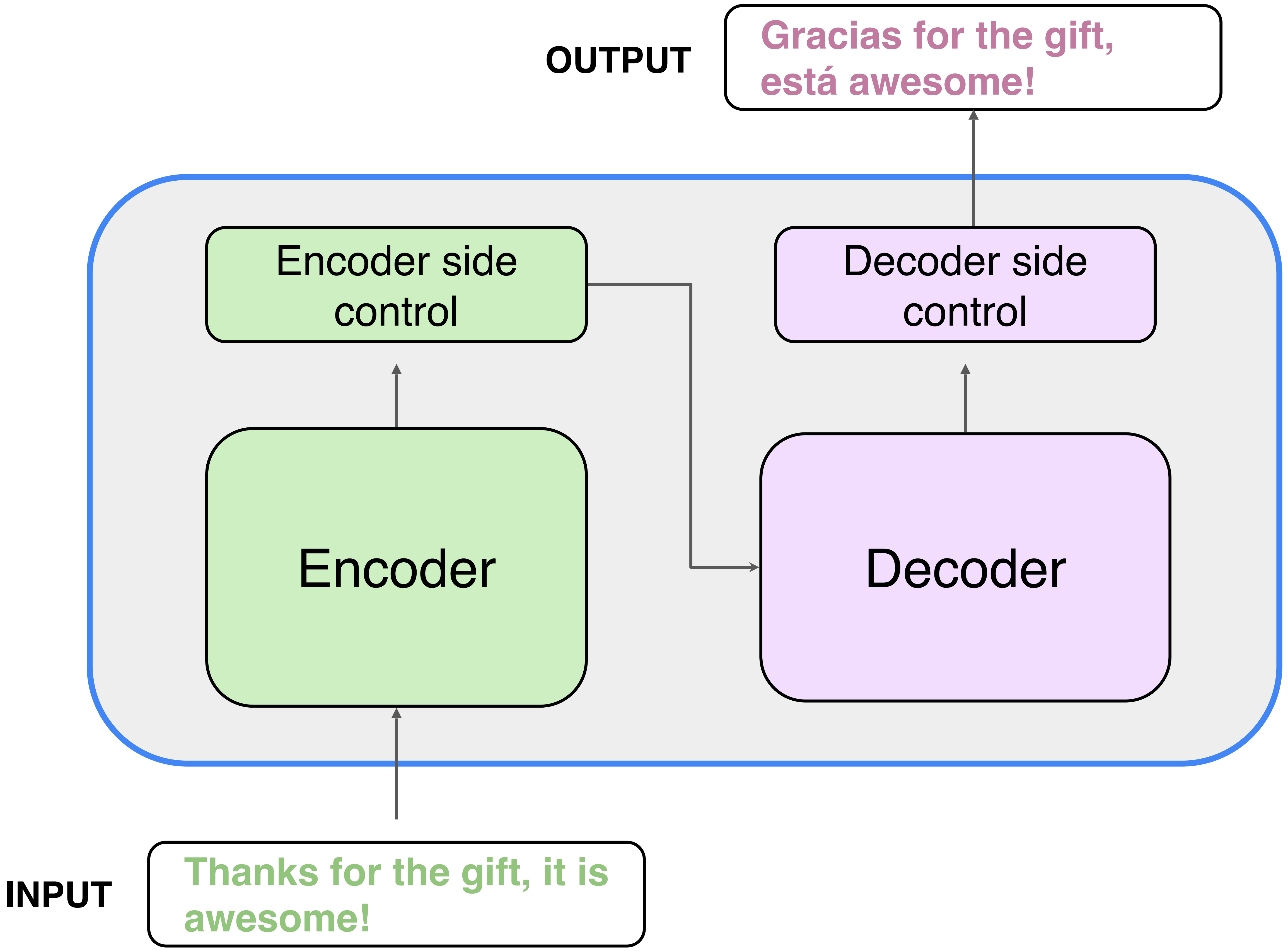
thee hai tomorrow 8 am ka alarm laga diya

okay alarm set kiya for 8 am

CoCoa: Controllable Code-switched Generation MRPJR'22

Modified sequence to sequence model

- **Control** attributes responsible for diversity at inference
  - Encoder side control
  - Decoder side control

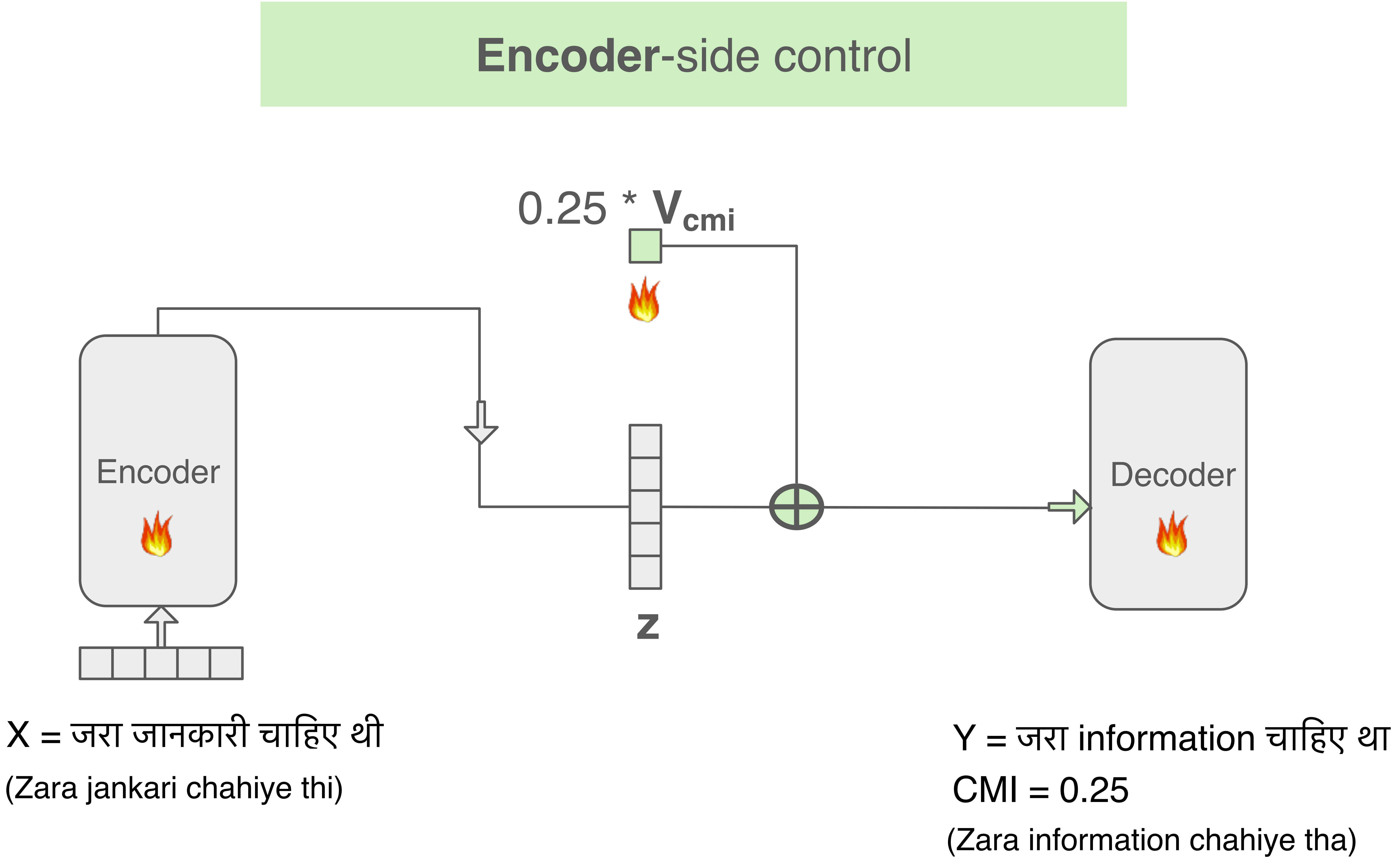


MRPJR'22 Mondal et al., "COCOA: An Encoder-Decoder Model for Controllable Code-switched Generation", EMNLP 2022

Encoder-side Control

- Used with attributes for which we have parallel monolingual to code-switched text with attribute values

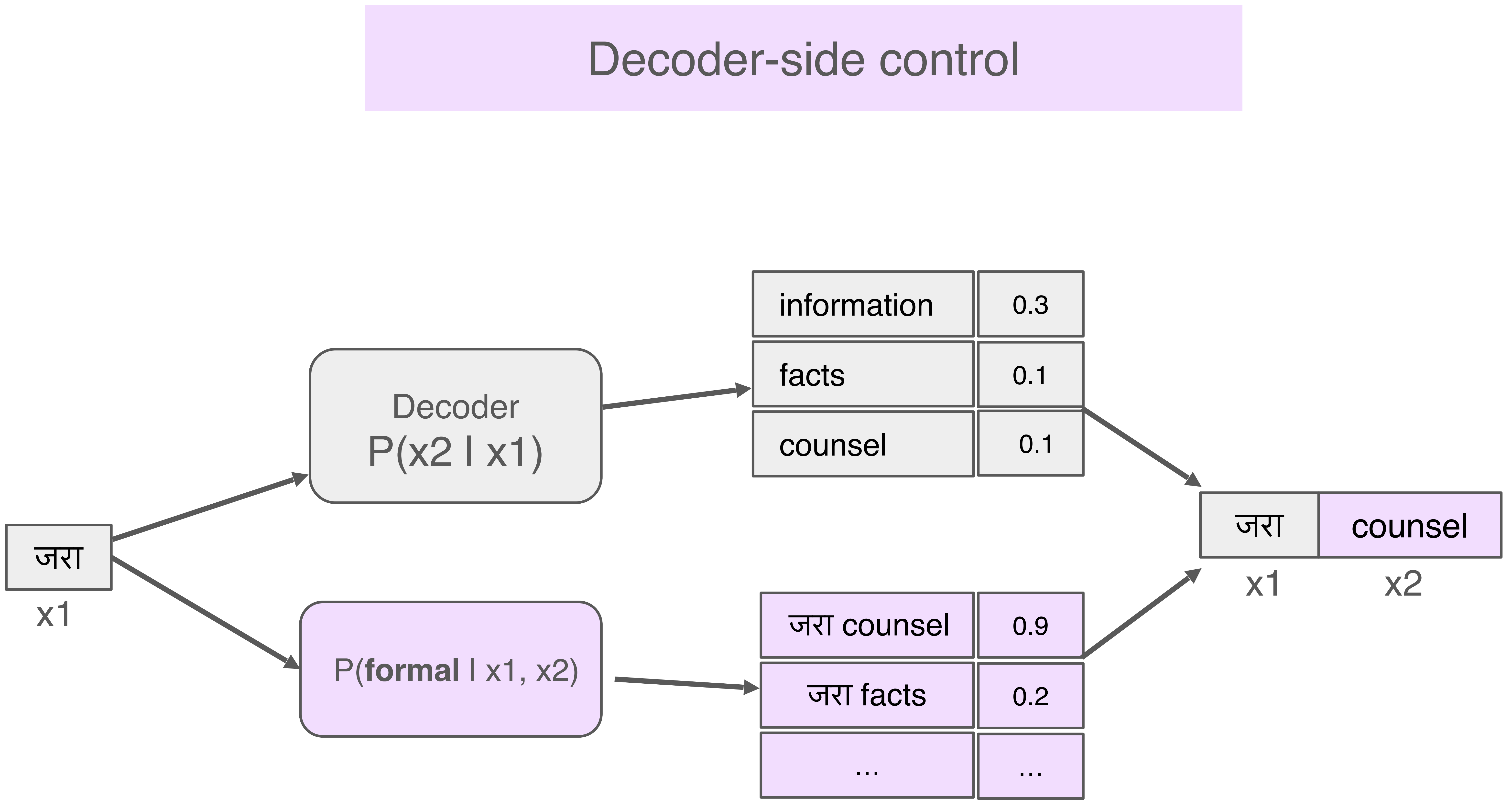
- Learn a vector embedding for each attribute, scale it with a weight (proportional to the attribute value) and add to the encoder representation SVSF'21



SVSF'21 "Controlling machine translation for multiple attributes with additive interventions",Schioppa et al., EMNLP 2021

Decoder-side Control

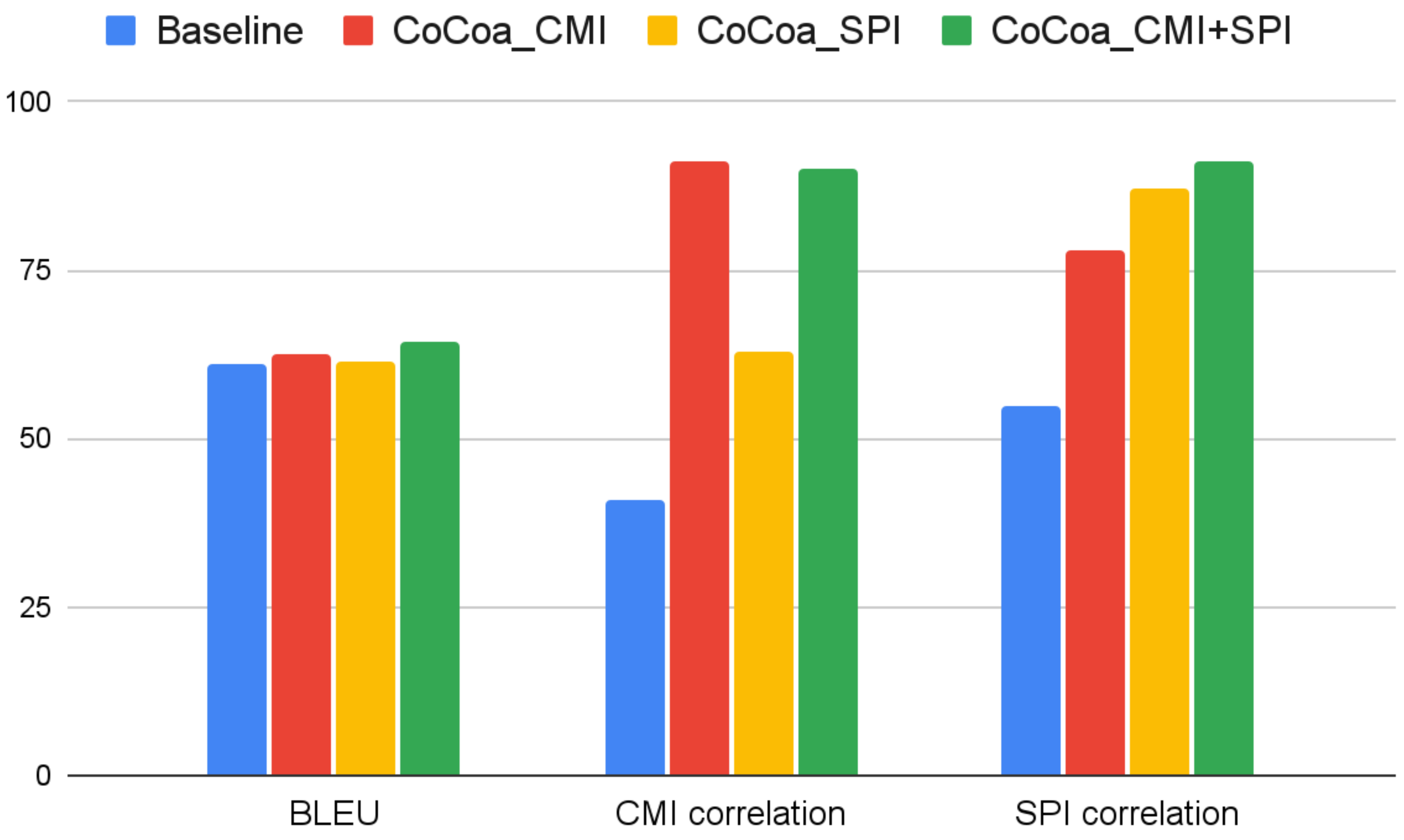
- Used with attributes, like formality, for which we do not have parallel text
- Predict using a binary attribute classifier whether each prefix string, on completion, will satisfy attribute or not YK’21
- Multiply probabilities from attribute classifier with output probability distributions and renormalize



YK’21 “FUDGE: Controlled text generation with future discriminators.”,Yang and Klein, NAACL 2

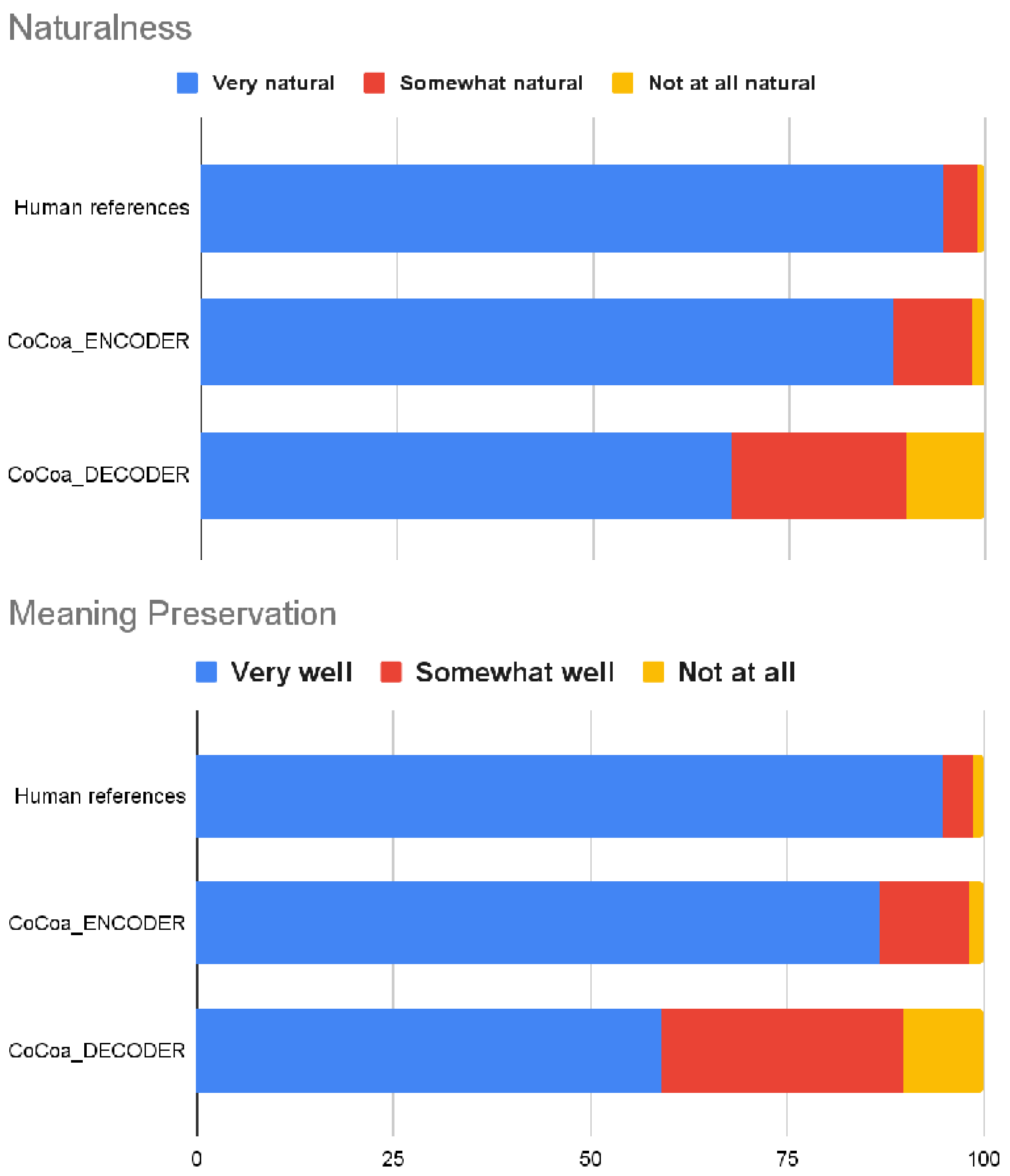
CoCoa: Generation Quality

- BLEU measures the quality of generated text
- Pearson correlation between attribute values of human references and model outputs



CoCoa: Human Evaluations

- Naturalness
- Meaning Preservation
  - Encoder-based control produces more natural and consistent outputs
  - Decoder-based control achieves attribute faithfulness at the cost of naturalness



CoCoa: Examples of Generations

Hindi: उसे भाग लेने की इजाजत नहीं थी

cmi-low: उसे भाग लेने की permission नहीं थी

cmi-high: उसे participate लेने की permission नहीं थी

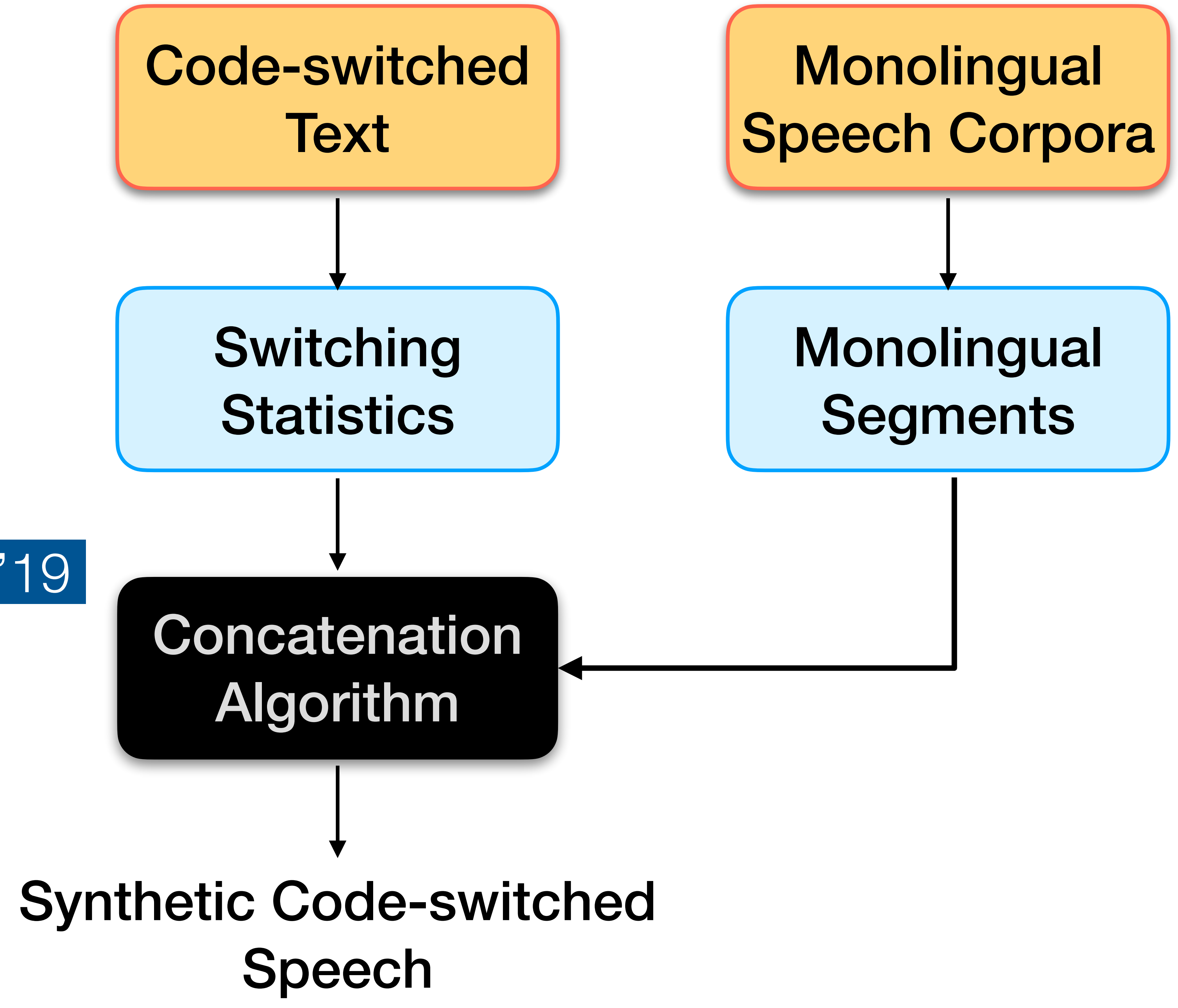
Hindi: उन्होंने मुझसे कहा की अंत में एक व्यक्ति से मीटिंग करनी होगी

cmi-low: उन्होंने मुझसे कहा की end में एक व्यक्ति से meeting करनी होगी

cmi-high: they told me की end में एक व्यक्ति से meeting करनी होगी

Synthesizing Code-switched Speech?

- Hard to access large amounts of code-switched data
- Can we leverage monolingual speech to construct synthetic code-switched speech?
  - Create synthetic speech that mimics phonetic constraints of real code-switched speech at switching boundaries TGJA'19
  - Can we use text-to-speech synthesis systems to generate synthetic code-switched speech? SATJ'20



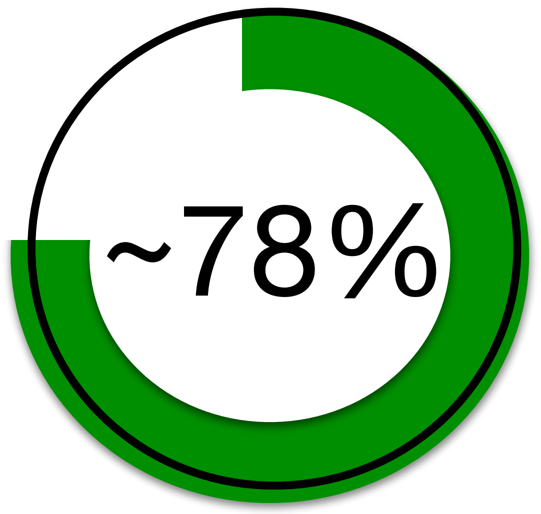
TGJA'19 "Exploiting Monolingual Speech Corpora for Code-mixed Speech Recognition", K. Taneja, S. Guha, P. Jyothi, B. Abraham, Interspeech 2019

SATJ'20 "Improving Low-resource Code-switched ASR using Augmented Code-switched TTS", Y. Sharma, B. Abraham, K. Taneja, P. Jyothi, Interspeech 2020

Summary

- ASR on accented speech from underrepresented users remains unsolved
- Code-switched inputs are still hard for computational models to proces

Critical to ensure more inclusive adoption of speech technologies



Google witnessed a whopping 78% jump in voice search from 2021 to 2022

Voice-based inputs make technology accessible to those who cannot type in their native languages

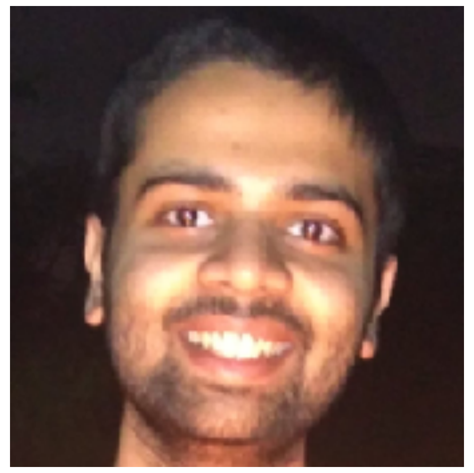
Thanks to all my collaborators!


Abhijeet Awasthi


Aman Kansal


Kartik Khandelwal


Archiki Prasad


Syamantak Kumar


Ishan Tarunesh


Ritika


Shreya Pathak


Vinit Unni


Sneha Mondal


Aravindan Raghuveer


Sunita Sarawagi