Speech Recognition for Under-resourced Languages using Probabilistic Transcriptions

Preethi Jyothi

Department of CSE, IIT Bombay

CS344 Guest Lecture

February 7, 2017

Introduction

Automatic speech recognition (ASR):

Translate spoken words into text

ASR isn't to blame for this...

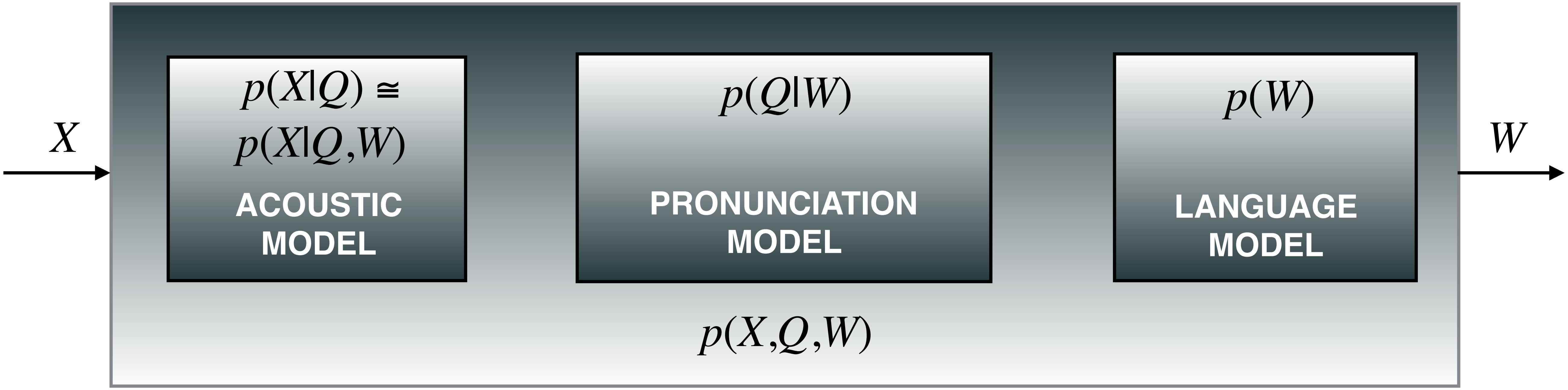
Image from http://takeitwithagrainofsalt.quora.com/Thanks-Siri?srid=pr0Y&share=1

Introduction

Automatic speech recognition (ASR):
Translate spoken words into text

Modern ASR systems are dominated by statistical methods pioneered by [Jelenik '76]

Standard ASR Pipeline



Decoding: Given X, find $\underset{W}{\operatorname{argmax}}\, p(W|X)$

$$= \underset{W}{\operatorname{argmax}}\, p(W, X)$$

$$= \underset{W}{\operatorname{argmax}} \sum_{Q} p(X,Q,W)$$

ASR over the years

- Great progress in ASR performance
  - Aided by algorithmic and computational advances
- Recently: Baidu's *Deep Speech 2* comparable to human performance
  - Trained on about 10,000 hours of labeled speech
- But limited language diversity

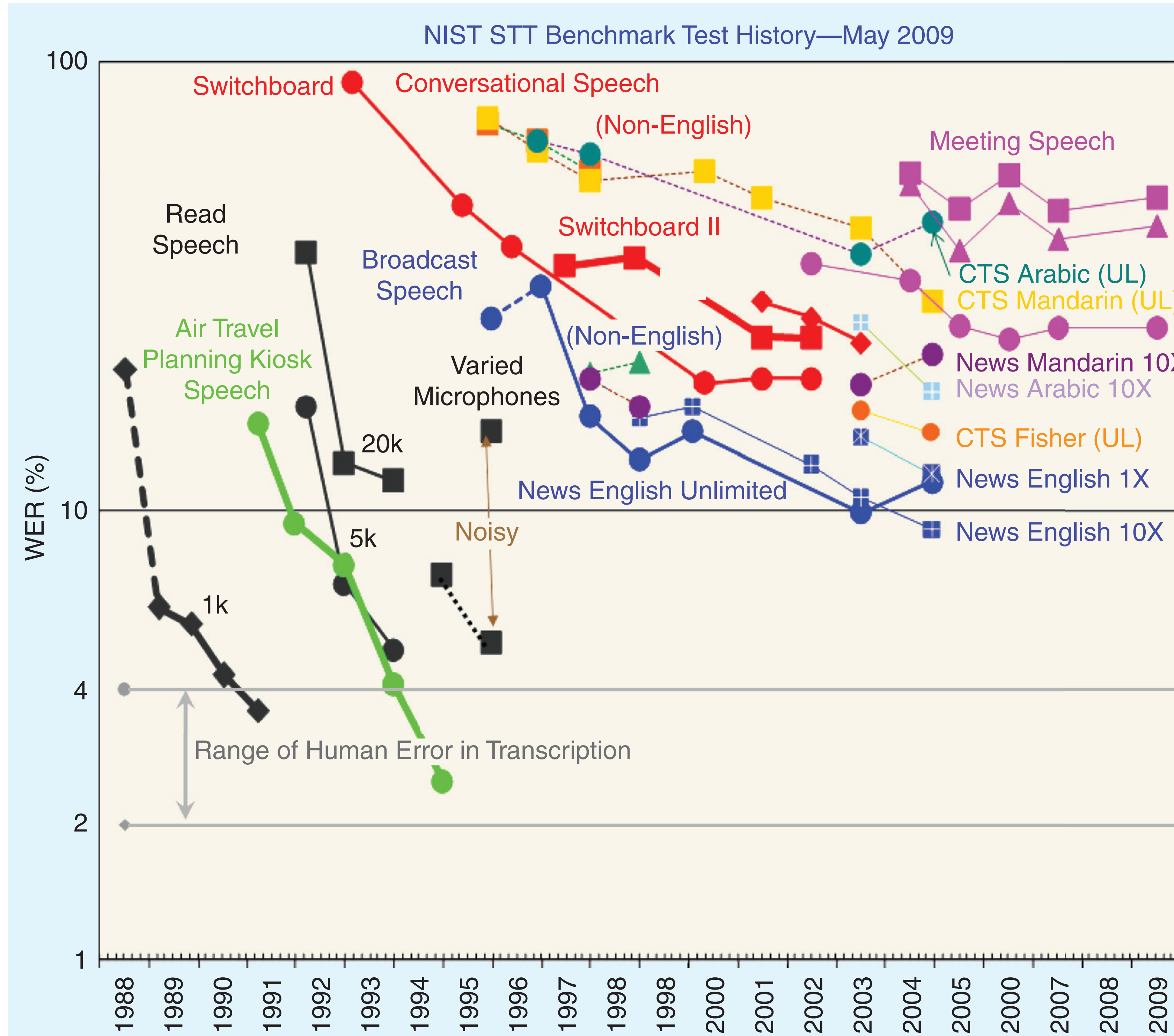
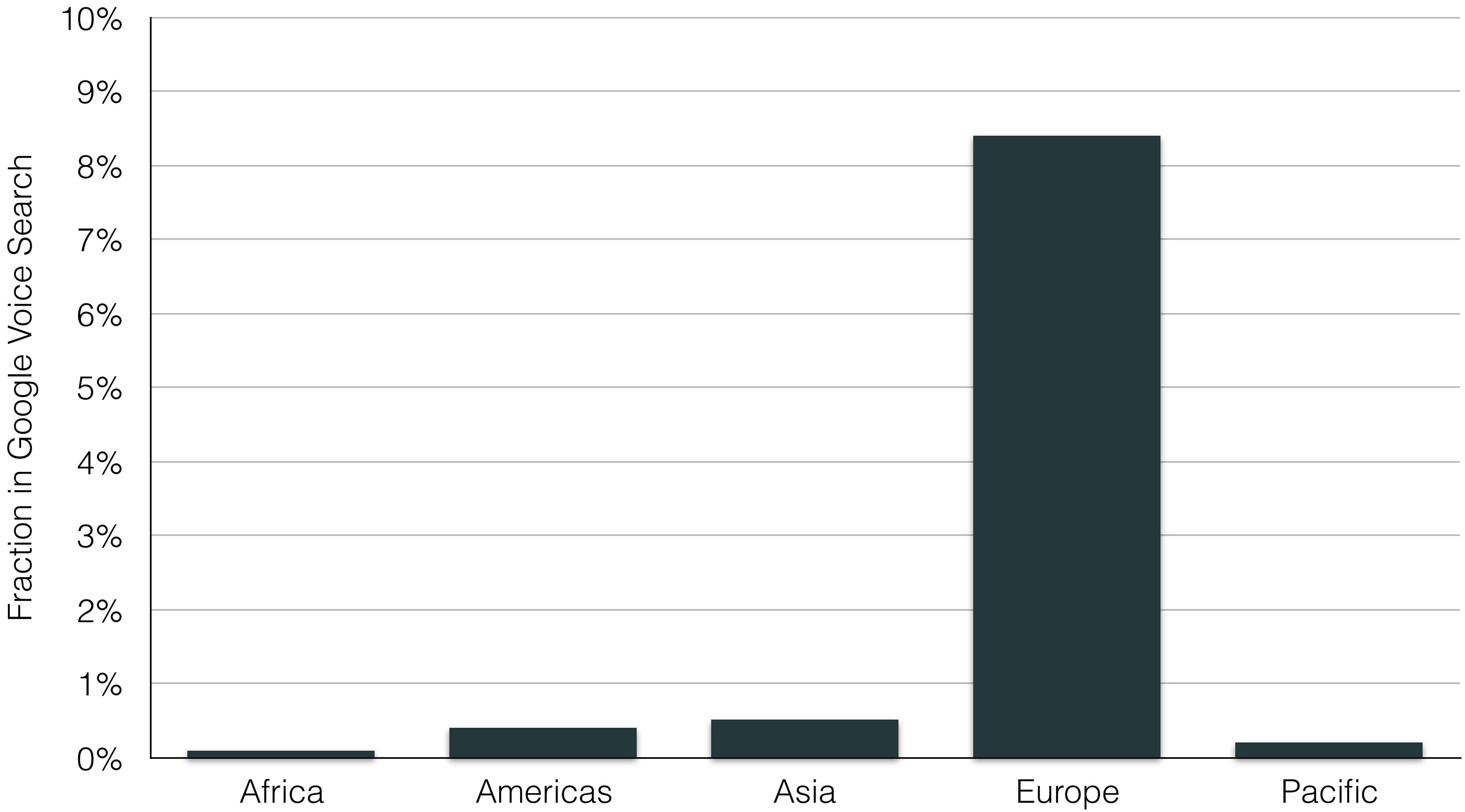


Image reproduced from He & Deng, *IEEE Signal Proc. Magazine*, 2011

Languages with ASR

E.g., Google Voice Search supports < 80 out of 7000 languages



https://googleblog.blogspot.com/2012/08/voice-search-arrives-in-13-new-languages.html

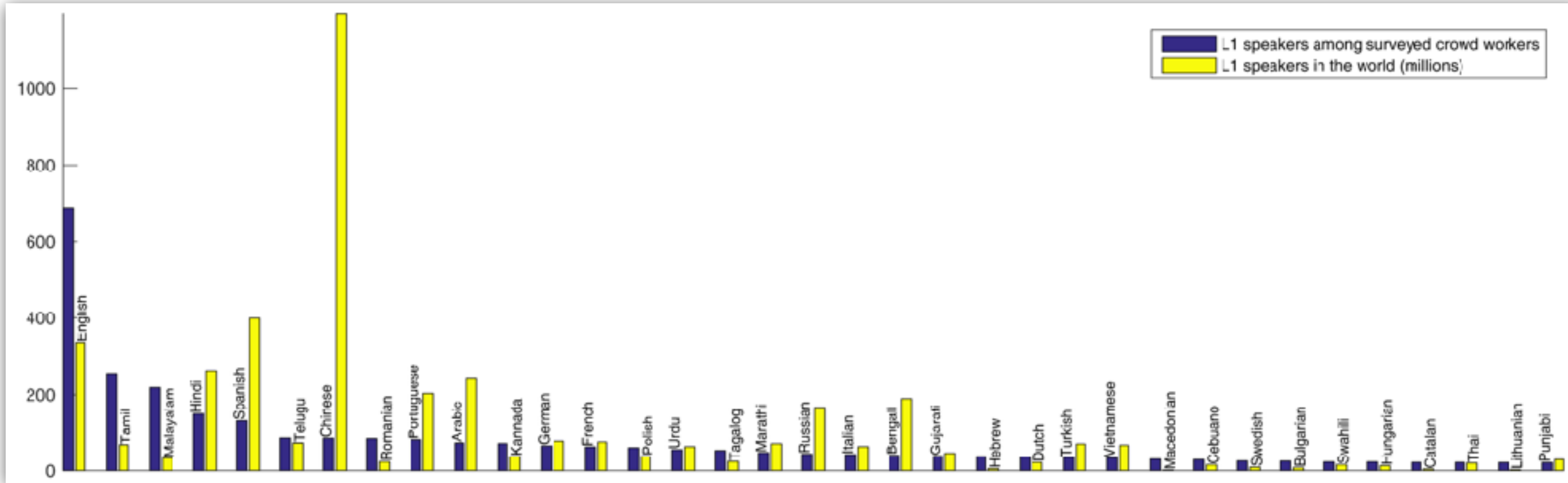
ASR for all languages

- ASR systems in all languages?
  - Speech is the primary means of human communication
  - Develop natural interfaces for both literate & illiterate users
  - Contribute to preservation of endangered languages

Lack of Transcribed Corpora

- Major challenge: Building ASR systems is very data-hungry
  - Require large amounts of labeled speech data: Speech audio with *matching transcriptions*
  - Transcription by native speakers is a laborious and expensive process
- Crowdsourcing might help alleviate the problem
  - However, *significant* mismatch in native languages of crowd workers and native language populations in the world

Native Language Mismatch



- Very few (to zero) crowd workers speak minority languages
- Distributional mismatch between language background of crowd workers with the language expertise required to complete transcription tasks

Mismatched Crowdsourcing

- A major bottleneck for ASR in new languages: Labeled speech
- Transcribers need to be native speakers

Use Non-native Speakers?

Mismatched Crowdsourcing

How can it possibly work?![1]

[1][Jyothi & Hasegawa-Johnson *AAAI*-15 & *Interspeech*-15]

Mismatched Crowdsourcing

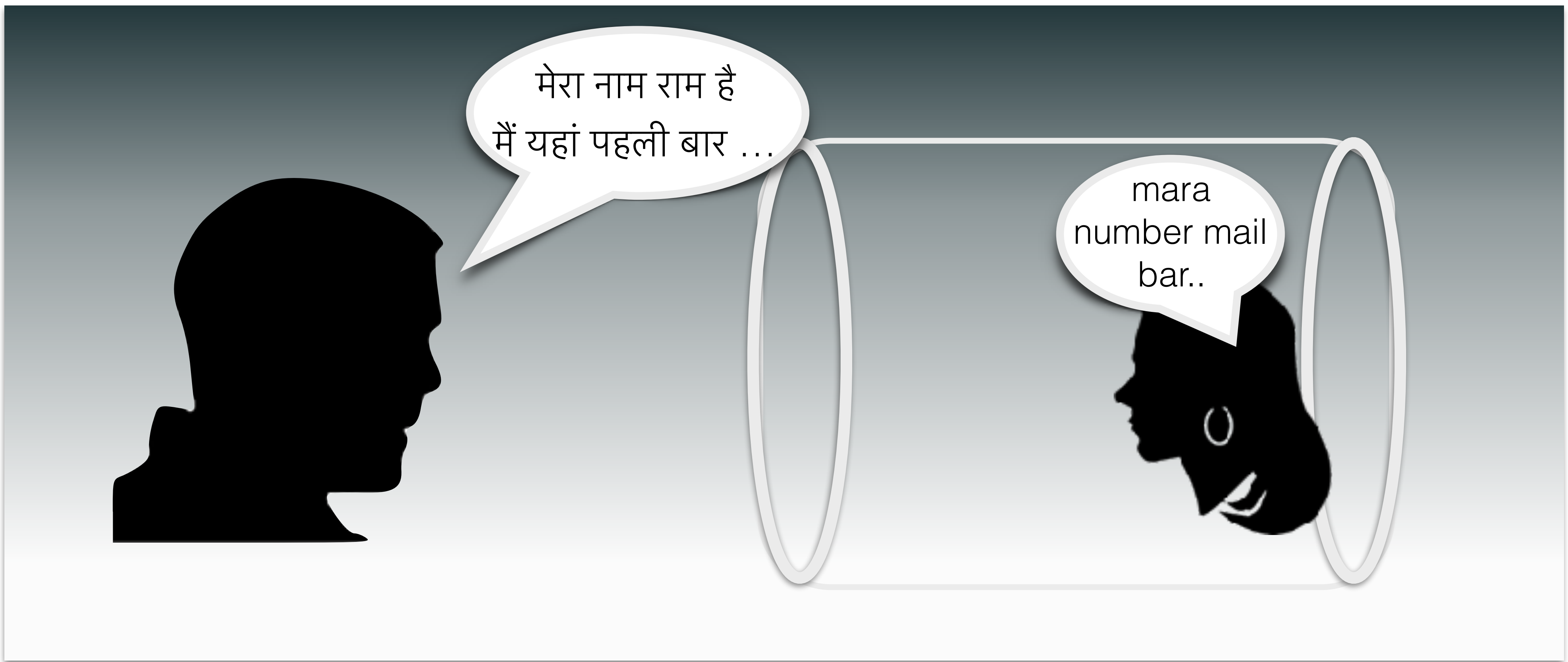
- How can it possibly work?! [Best '94, Flege '95, etc.]
  - We are typically bad at perceiving speech in foreign languages!
    - Unfamiliar sounds, no vocabulary, no language model to go by, distorted by native languages, …

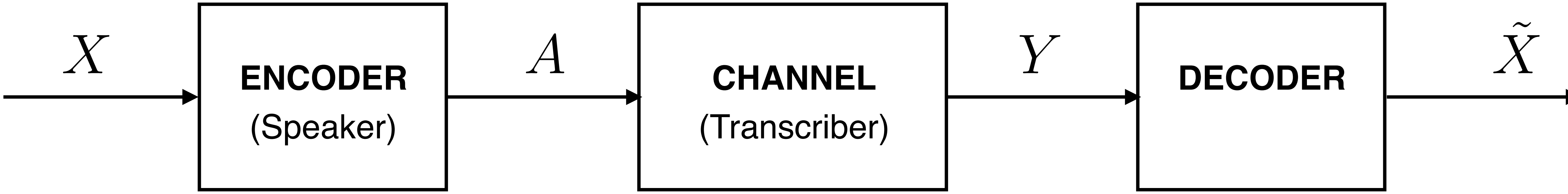
Mismatched Crowdsourcing

- How can it possibly work?! [Best ’94, Flege ’95, etc.]
  - We are typically bad at perceiving speech in foreign languages!
    - Unfamiliar sounds, no vocabulary, no language model to go by, distorted by native languages, …
  - An extremely noisy channel

Mismatched Crowdsourcing

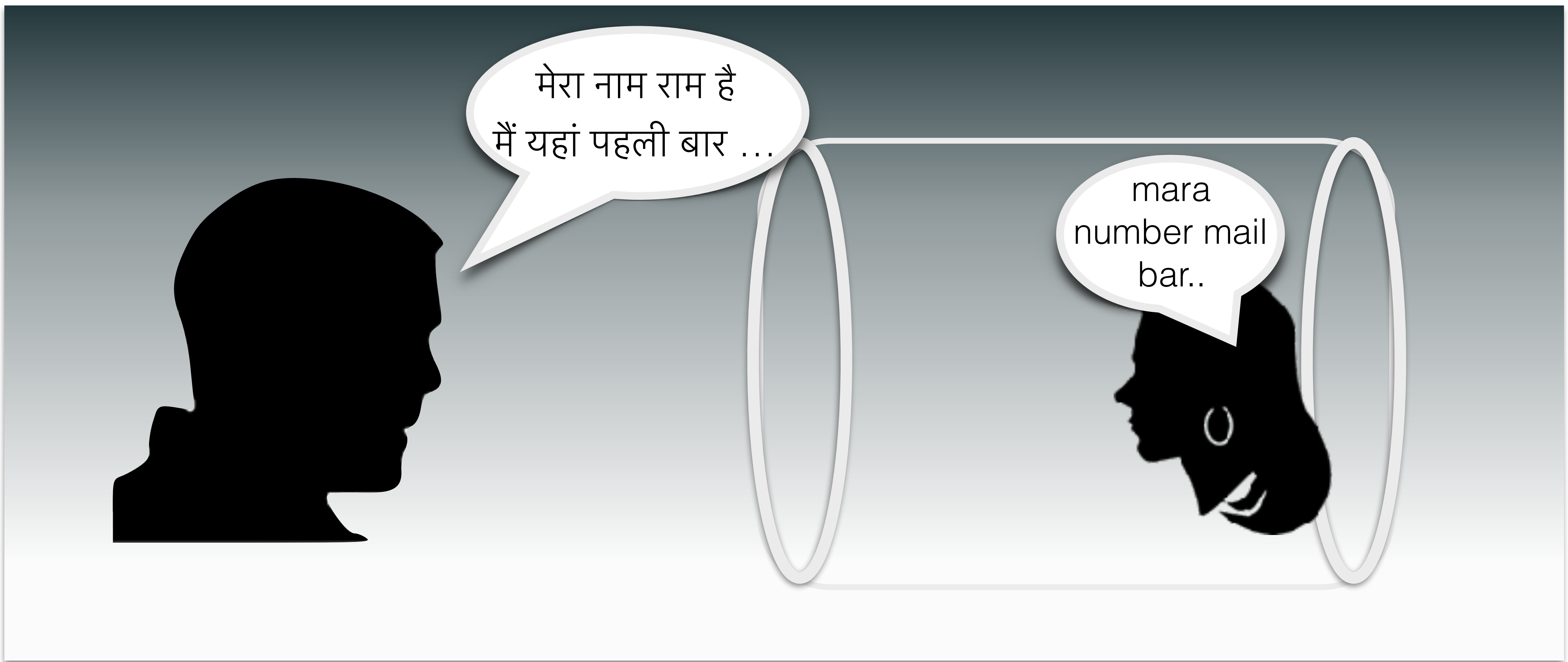
- An extremely noisy channel

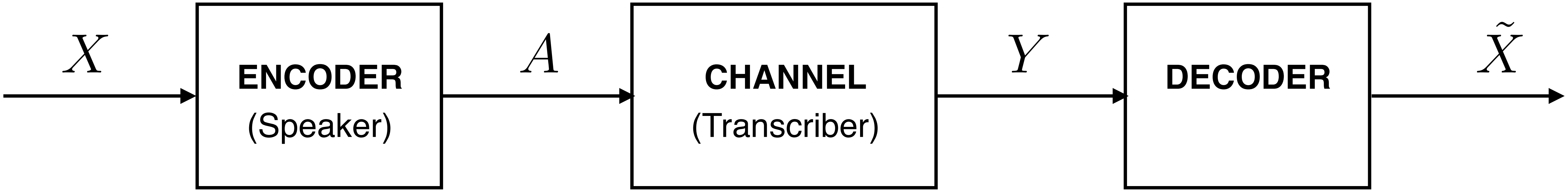




Solution: Error Correction

- Learn channel characteristics of the foreign listener

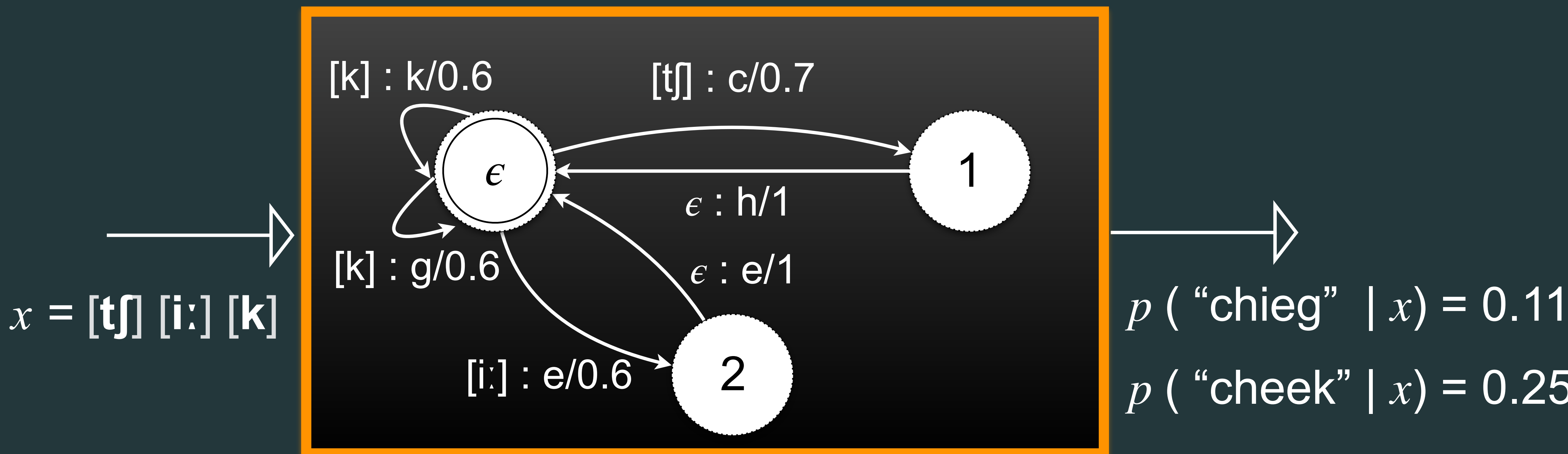




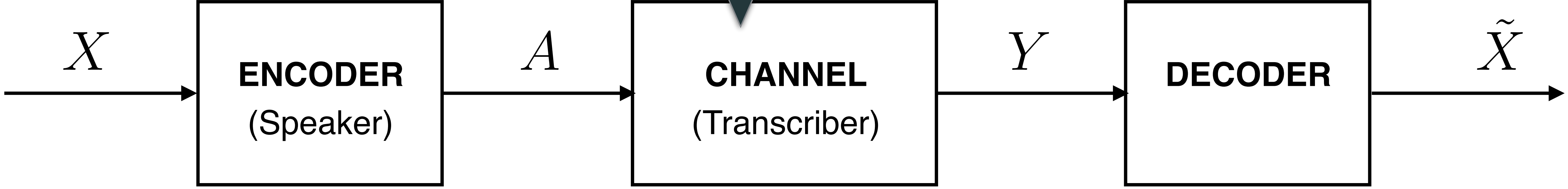
Solution: Error Correction

- Learn channel characteristics of the foreign listener

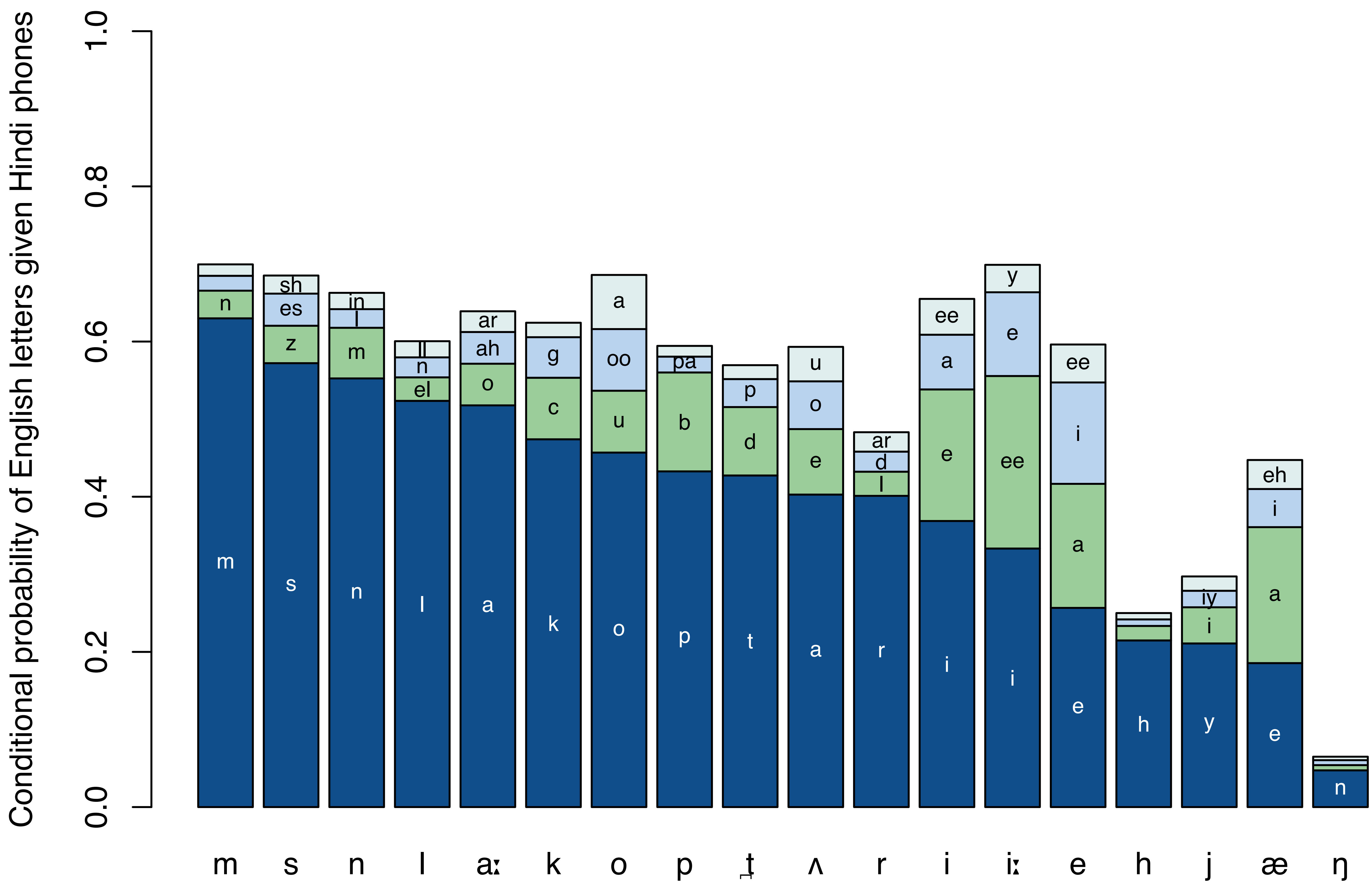
A Probabilistic Finite State model



Trained using the Expectation-Maximization (EM) algorithm

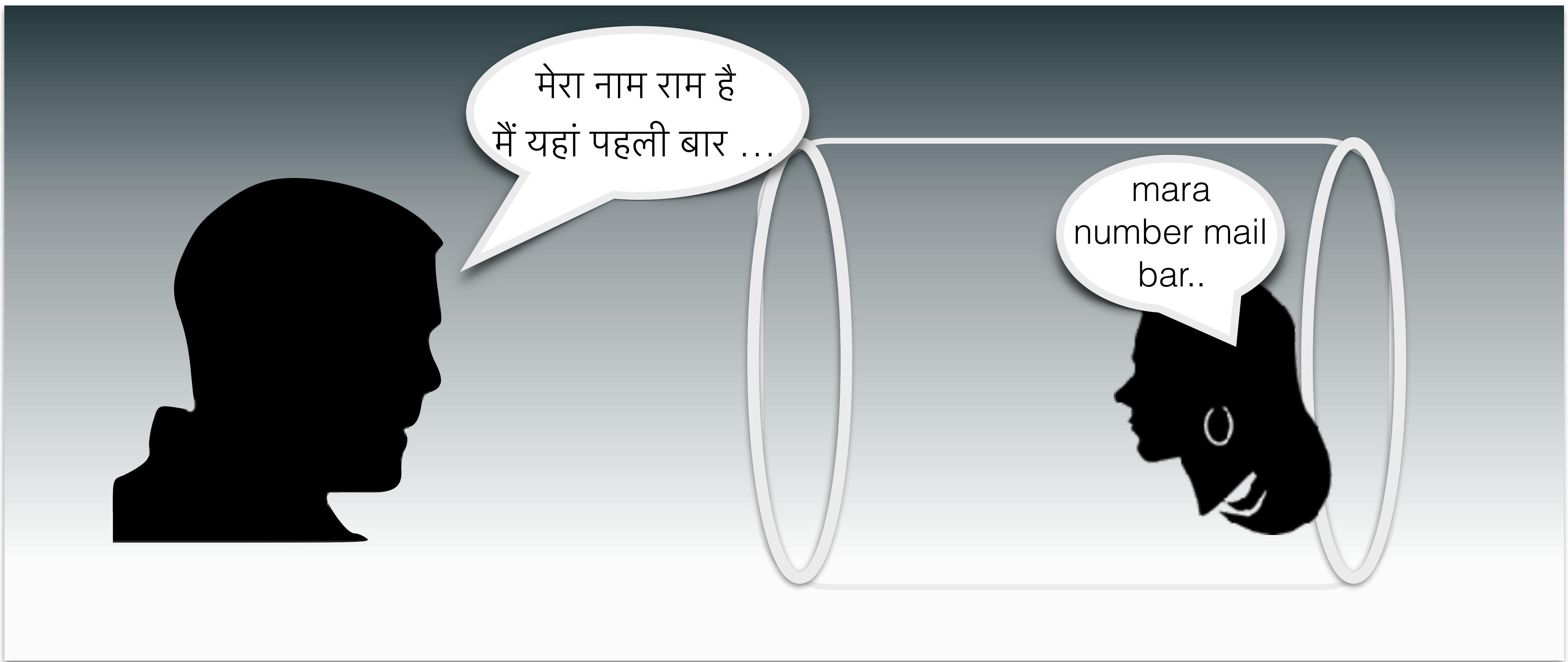


Visualizing the Channel



Solution: Error Correction

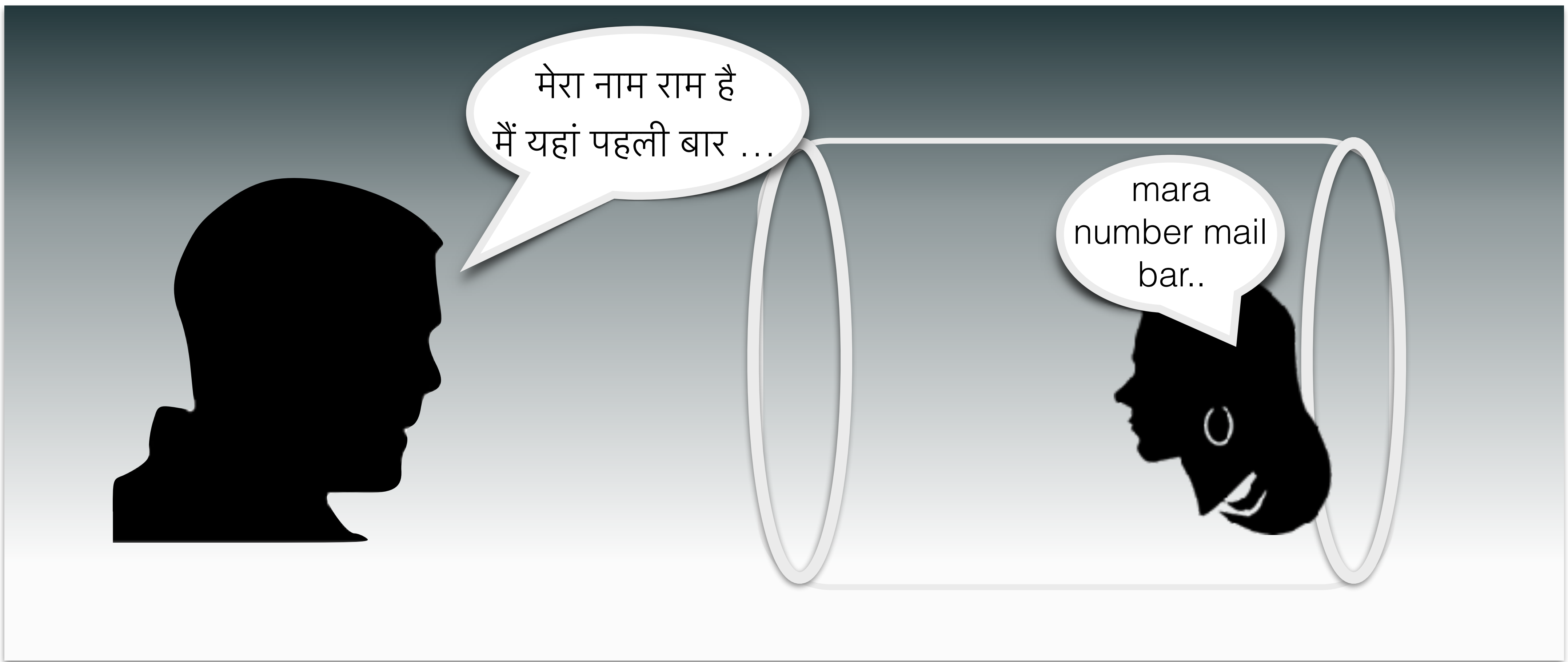
- Learn channel characteristics of the foreign listener
- But also need to use an error-correcting code

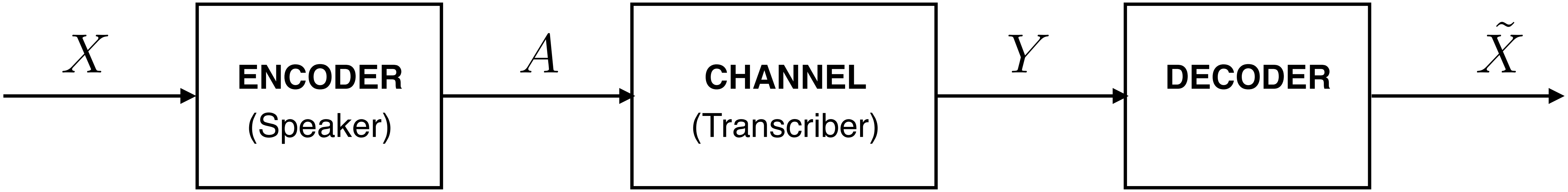


But how can we encode speech?

Solution: Error Correction

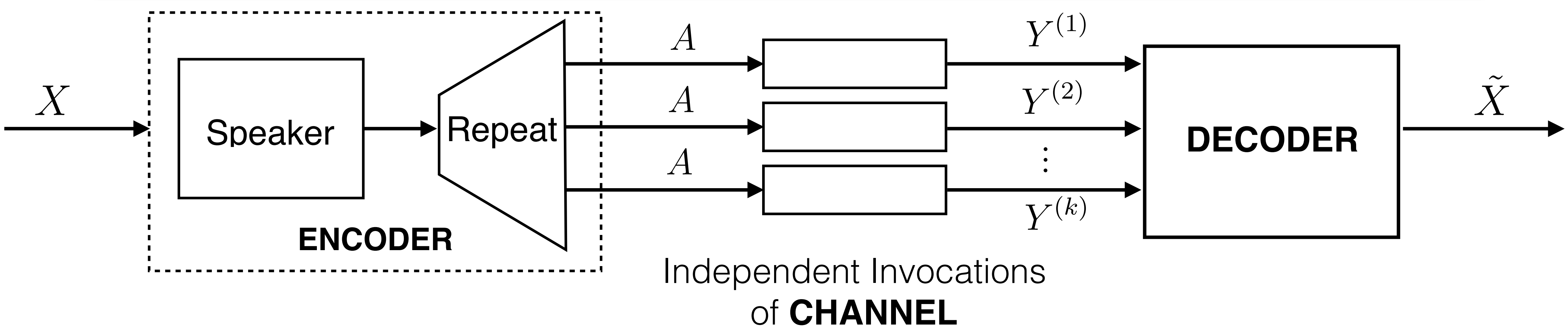
- Learn channel characteristics of the foreign listener
- Use a repetition code!



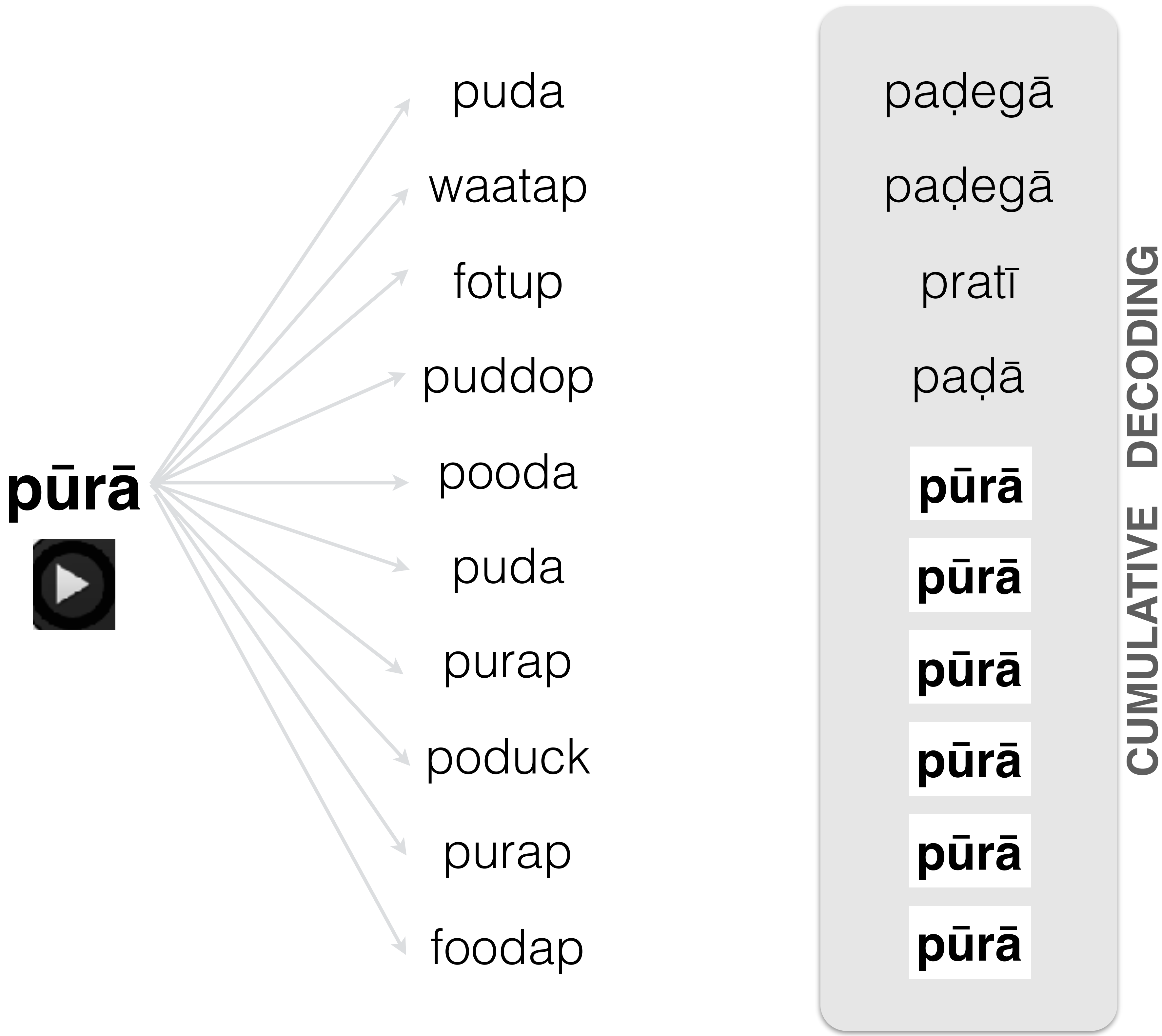


Solution: Error Correction

- Learn channel characteristics of the foreign listener
- Use a repetition code!



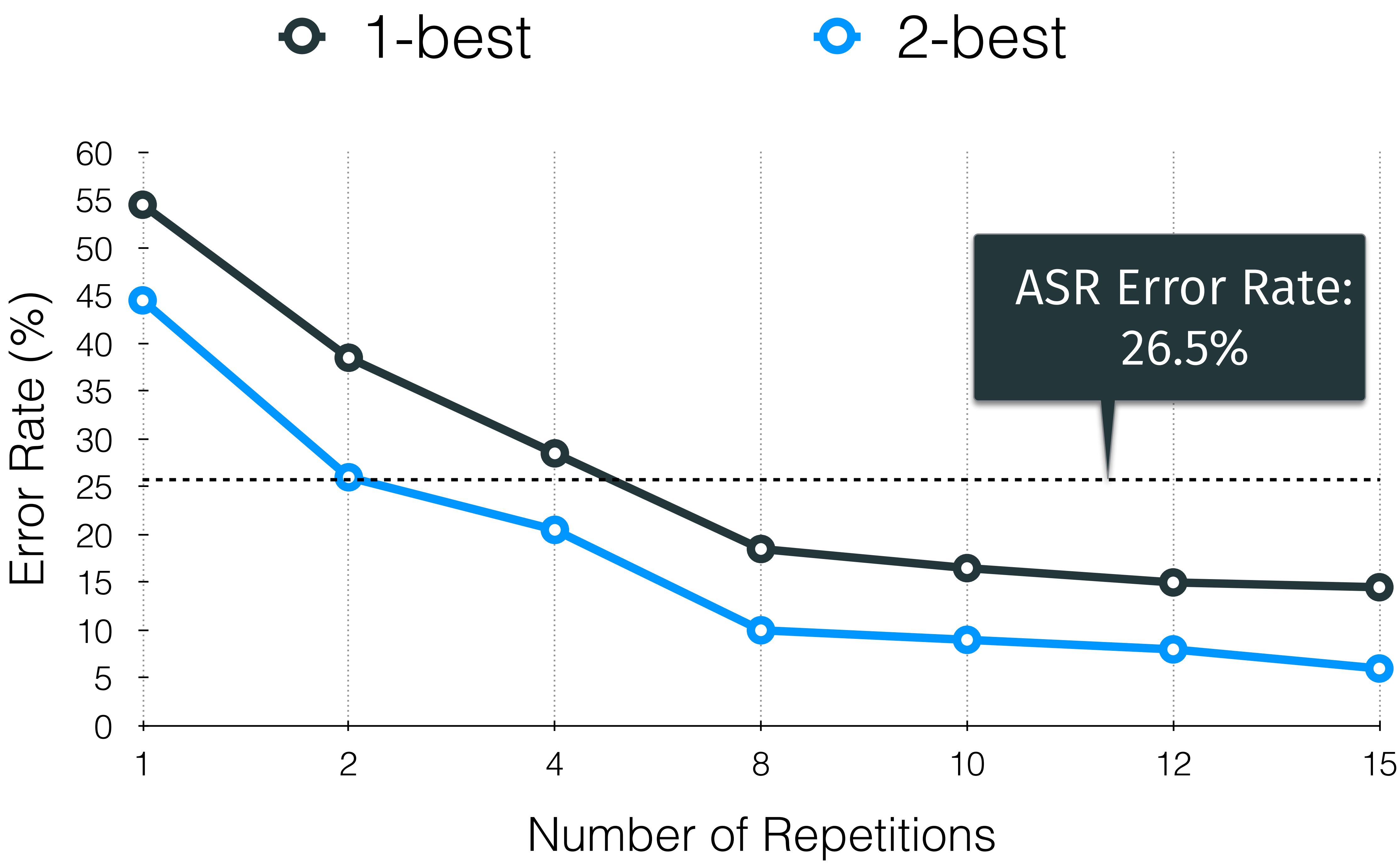
Example



pūrā
puda
waatap
fotup
puddop
pooda
puda
purap
poduck
purap
foodap
paḍegā
paḍegā
pratī
paḍā
pūrā
pūrā
pūrā
pūrā
pūrā
pūrā
CUMULATIVE DECODING

Labeling Error Rates

- Impressive accuracy (~5% error) on a medium-vocabulary isolated-word task!



[Jyothi & Hasegawa-Johnson *AAAI*-15]

Information-theoretic Analysis

- Conditional entropy, $H(X \mid Y)$, of the spoken words (Hindi), X, given the crowd transcripts Y, captures the amount of information lost in transmission
- $H(X \mid Y)$ can be naively upper-bounded using corpus cross-entropy
- However, errors in our channel model accumulate with increase in the number of repetitions, resulting in this upper-bound becoming less tight

Information-theoretic Analysis

Tighter bound on $H(X \mid Y)$ using an auxiliary random variable, $Z \in \{0,1\}$

Consider $W = \epsilon$ when $Z = 0$, $W = X$ when $Z = 1$

We set $Z = 1$ when $q(x/y)$ is sufficiently low

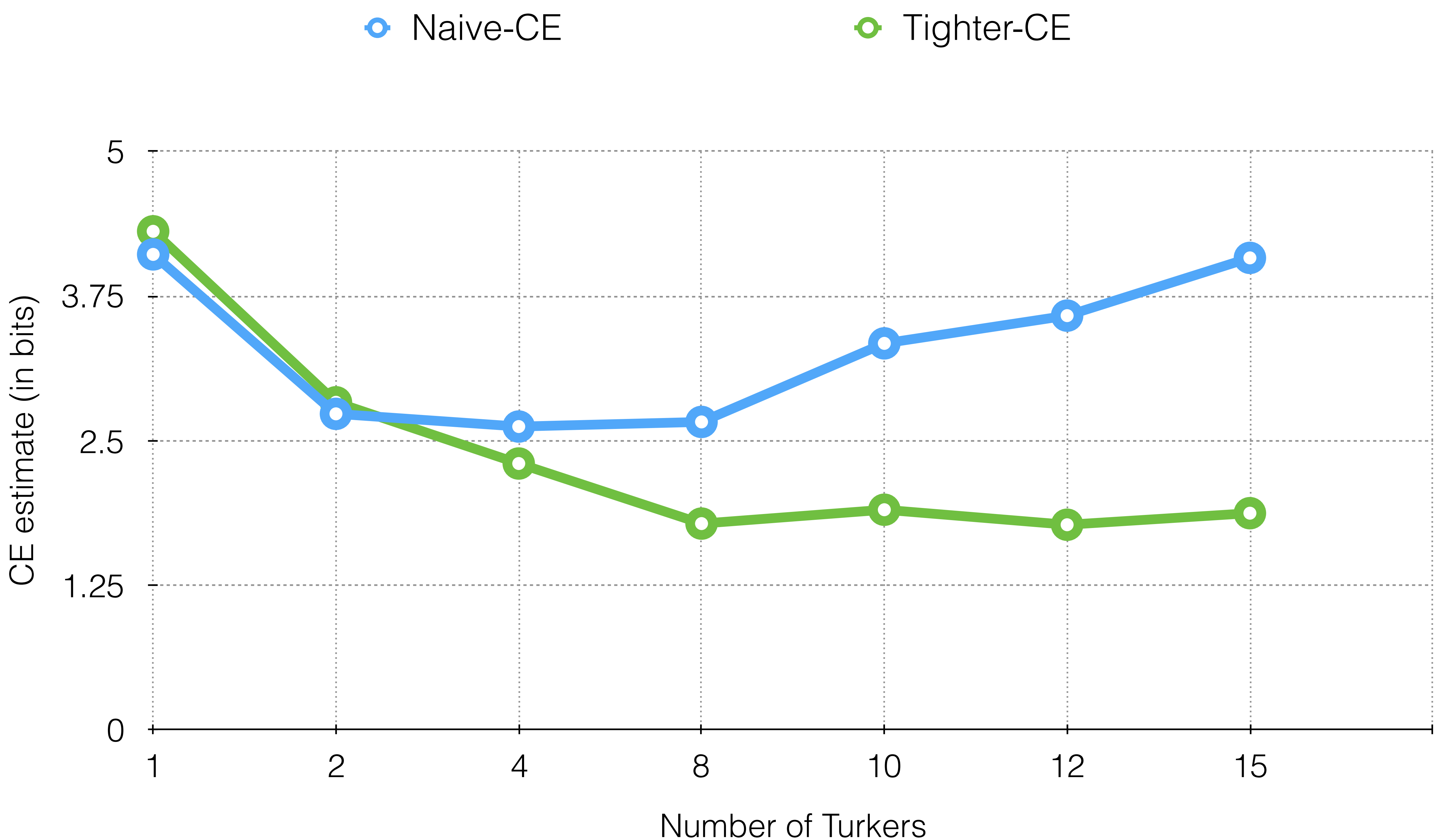
W represents side channel information indicating when the model needs to be corrected

$$H(X \mid Y) \leq p_0 \cdot \boxed{H(X \mid Y, Z{=}0)} + H(Z) + (1 - p_0) \log |\mathsf{X}|$$

where $p_0 = p(Z{=}0)$ and X is the input alphabet

Upper-bounded using corpus cross-entropy

Conditional Entropy Estimates

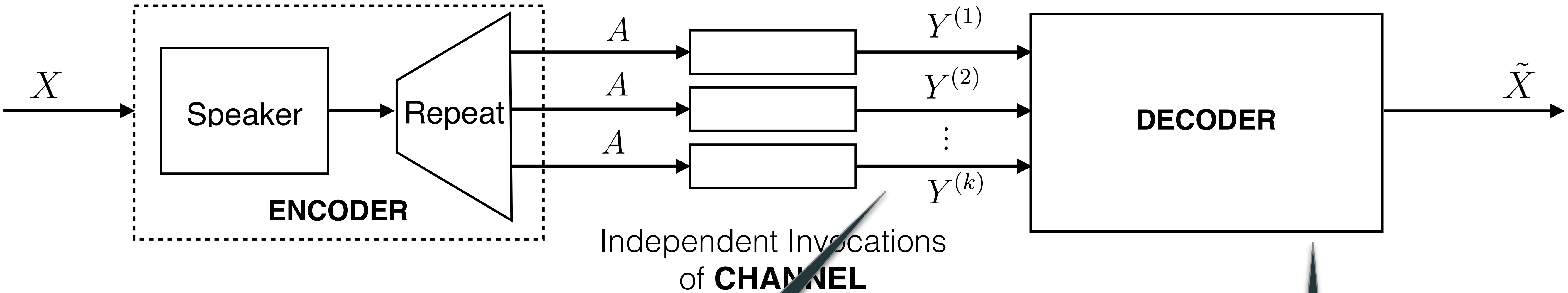


Our upper-bound estimates for CE are clearly tighter than the naive cross-entropy upper-bound estimate

Mismatched Crowdsourcing for Continuous Speech

February 7, 2017

Continuous Speech

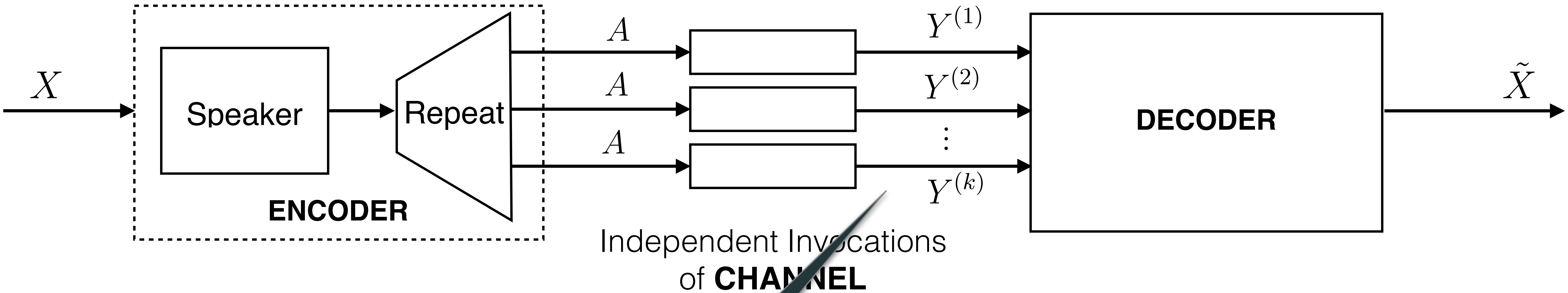


Decode each transcript individually using Maximum Likelihood Decoding and pick the output with the best score

Exact Maximum-Likelihood Decoding of multiple strings is intractable for long utterances

Word error rate: 77%

Continuous Speech



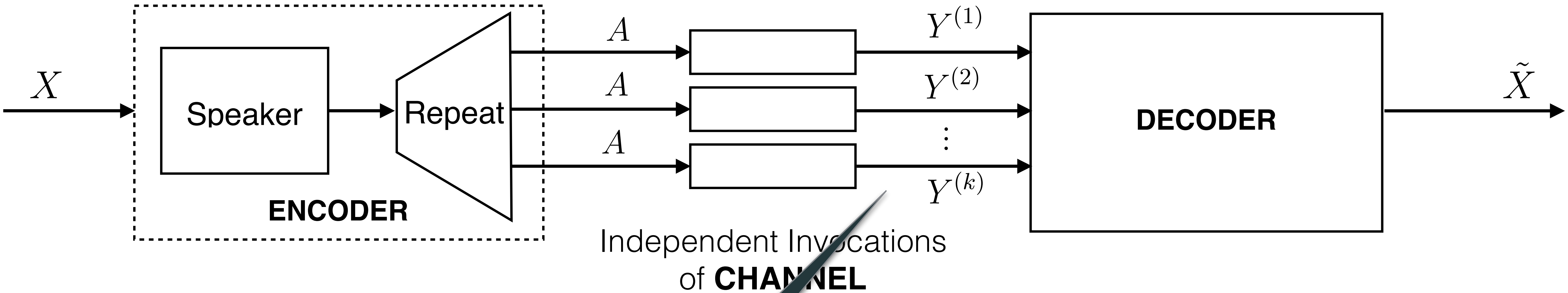
Decode each transcript individually using Maximum Likelihood Decoding and pick the output with the best score

Word error rate: 77%

Can we do better?

An outlier with a good score shouldn't be chosen over what many similar looking transcripts predict

Continuous Speech

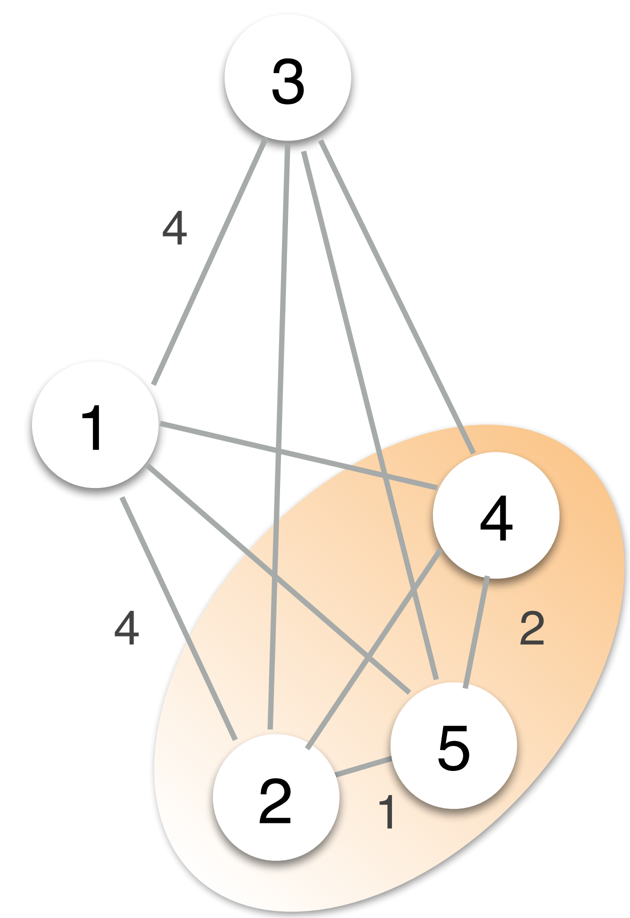


<u>Data Filtering:</u>
Use an edit-distance based similarity metric to discover a “cluster” to retain.

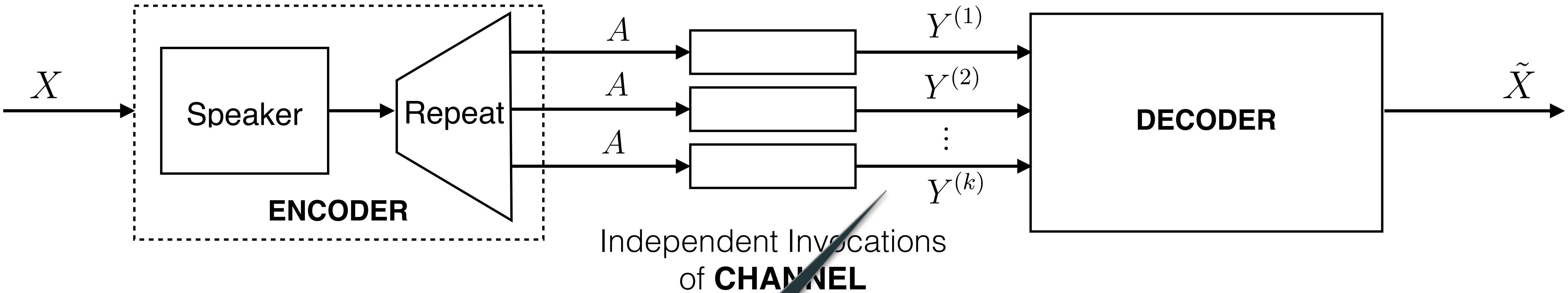
<u>Can we do better?</u>
An outlier with a good score shouldn’t be chosen over what many similar looking transcripts predict

Data Filtering

1	a	v	i	e	m	e	k	e
2	a	b	i	a	n	-	k	e
3	a	v	e	a	m	e	g	i
4	a	v	e	a	n	-	k	a
5	a	b	i	a	n	-	k	a



Continuous Speech

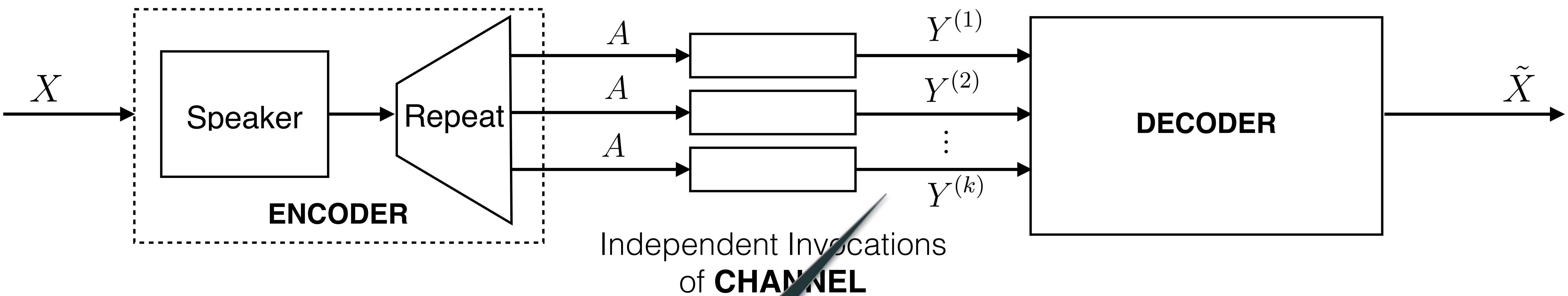


Data Filtering:
Use an edit-distance based similarity metric to discover a "cluster" to retain.

Word error rate: 68%

Can we do better?
An outlier with a good score shouldn't be chosen over what many similar looking transcripts predict

Continuous Speech

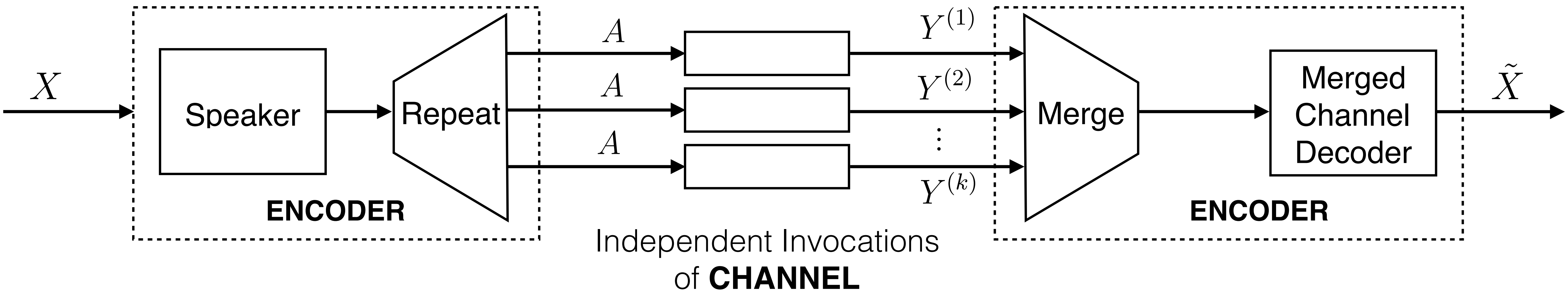


<u>Data Filtering:</u>
Use an edit-distance based similarity metric to discover a "cluster" to retain.

Word error rate: 68%

<u>Can we do even better?</u>

Channel Merger



Data Filtering

Discover "typical" transcripts

Alignment

NP-hard!

Approximation via incremental alignment algorithm

Alignment

k e e a j a g a
g i y a j a y g a
k e e a j a y g a h
c h a i j e g a

k	e	e	-	a	-	j	a	-	g	a	-
-	g	i	y	a	-	j	a	y	g	a	-
k	e	e	-	a	-	j	a	y	g	a	h
-	-	c	h	a	i	j	e	-	g	a	-

kiyā jāyegā

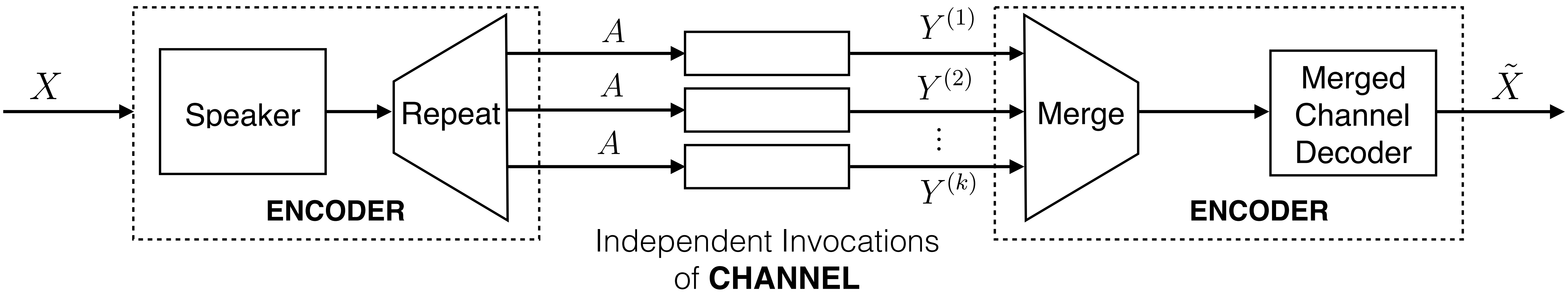
Alignment

k e e a j a g a
g i y a j a y g a
k e e a j a y g a h
c h a i j e g a

k	E	-	a	j	a	g	a	-
g	i	y	a	j	Y	g	a	-
k	E	-	a	j	Y	g	a	h
C	-	-	Y	j	e	g	a	-

kiyā jāyegā

Channel Merger



Data Filtering

Discover "typical" transcripts

Alignment

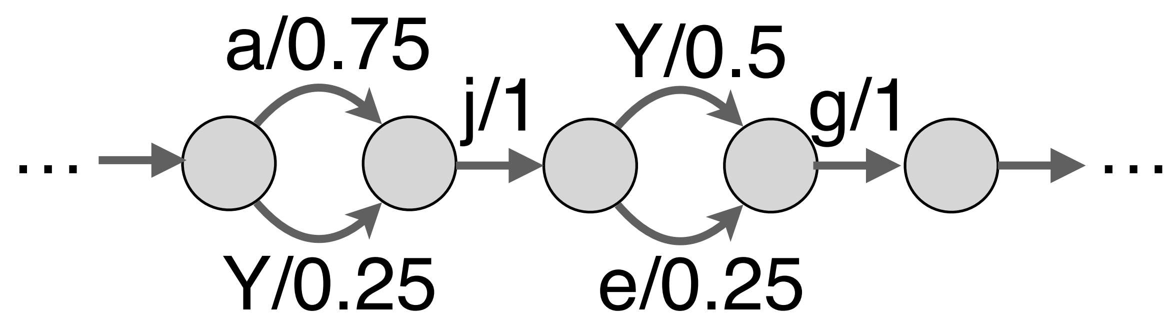
NP-hard!

Approximation via incremental alignment algorithm

Merge

Merge into one probabilistic transcript

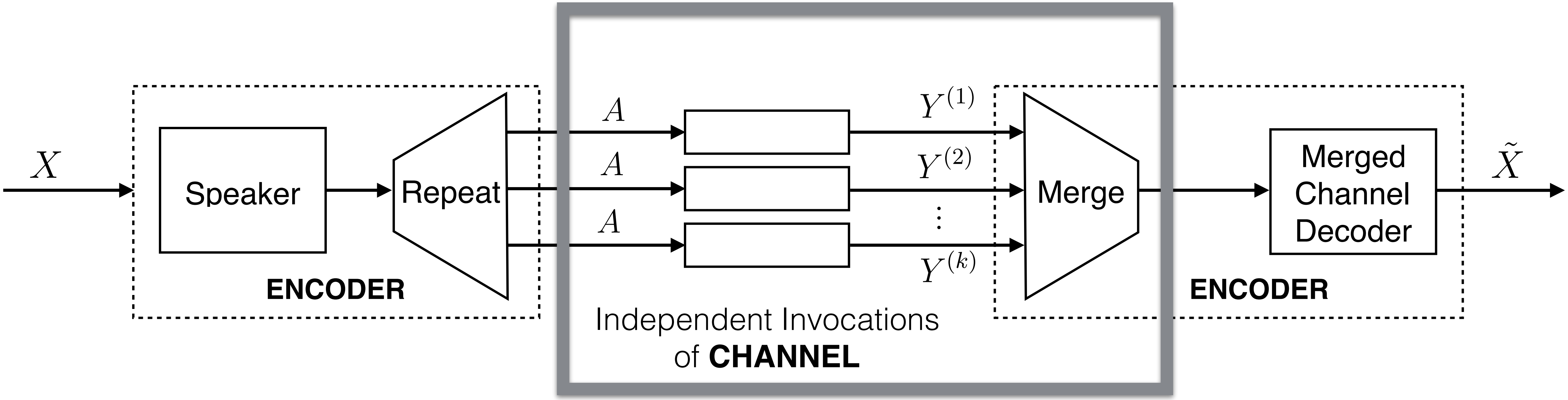
Merge Transcripts



k	E	-	a	j	a	g	a	-
g	i	y	a	j	Y	g	a	-
k	E	-	a	j	Y	g	a	h
C	-	-	Y	j	e	g	a	-

kiyā jāyegā

Channel Merger



Data Filtering

Discover "typical" transcripts

Alignment

NP-hard!

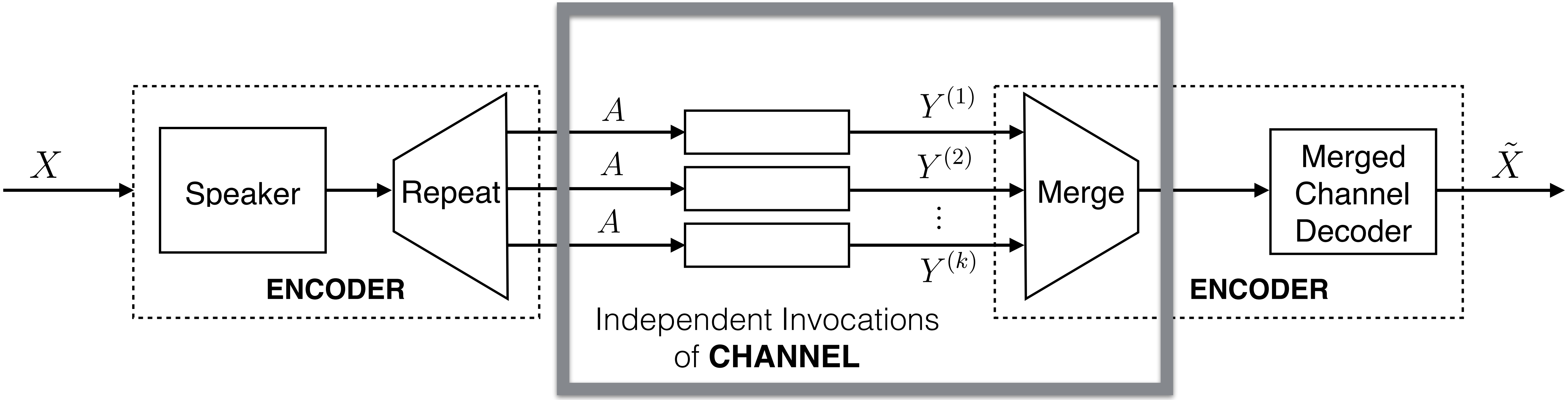
Approximation via incremental alignment algorithm

Merge

Merge into one probabilistic transcript

Model for merged channel

Channel Merger



Data Filtering

Discover "typical" transcripts

Alignment

NP-hard!

Approximation via incremental alignment algorithm

Merge

Merge into one probabilistic transcript

Model for merged channel

Shortlist & Decode

List Decoding + Exact Decoding from List

Probabilistic Transcriptions

```
Tacapo piza
strucka po zapecham
trakapo trabiza
Straka pose ta peesome
straka po ta pisha
strah kah poh chah peesh um
chaka-pu shapisha
stakkappoo sabeesham
takapo chapiser
Strike a pose some pizza
```

[1] [P. Jyothi & Hasegawa-Johnson, *Interspeech*-15]

Probabilistic Transcriptions

Tacapo piza

s	t	r	a	-	k	a	p	o	z	t	a	-	p	E	-	s	o	m
s	t	r	a	-	k	a	p	o	-	t	a	-	p	i	-	S	a	-
s	t	r	a	h	k	a	p	o	-	C	a	h	p	E	-	S	u	m
s	t	r	Y	-	k	a	p	o	z	s	a	m	p	E	t	s	a	-
s	t^h	r	ɑ/ ɑi		k^h	ɑ	p^h	ɔ		$tʃ^h$/ t/s	ɑ		p^h	i/i:		ʃ/s	ɑ	

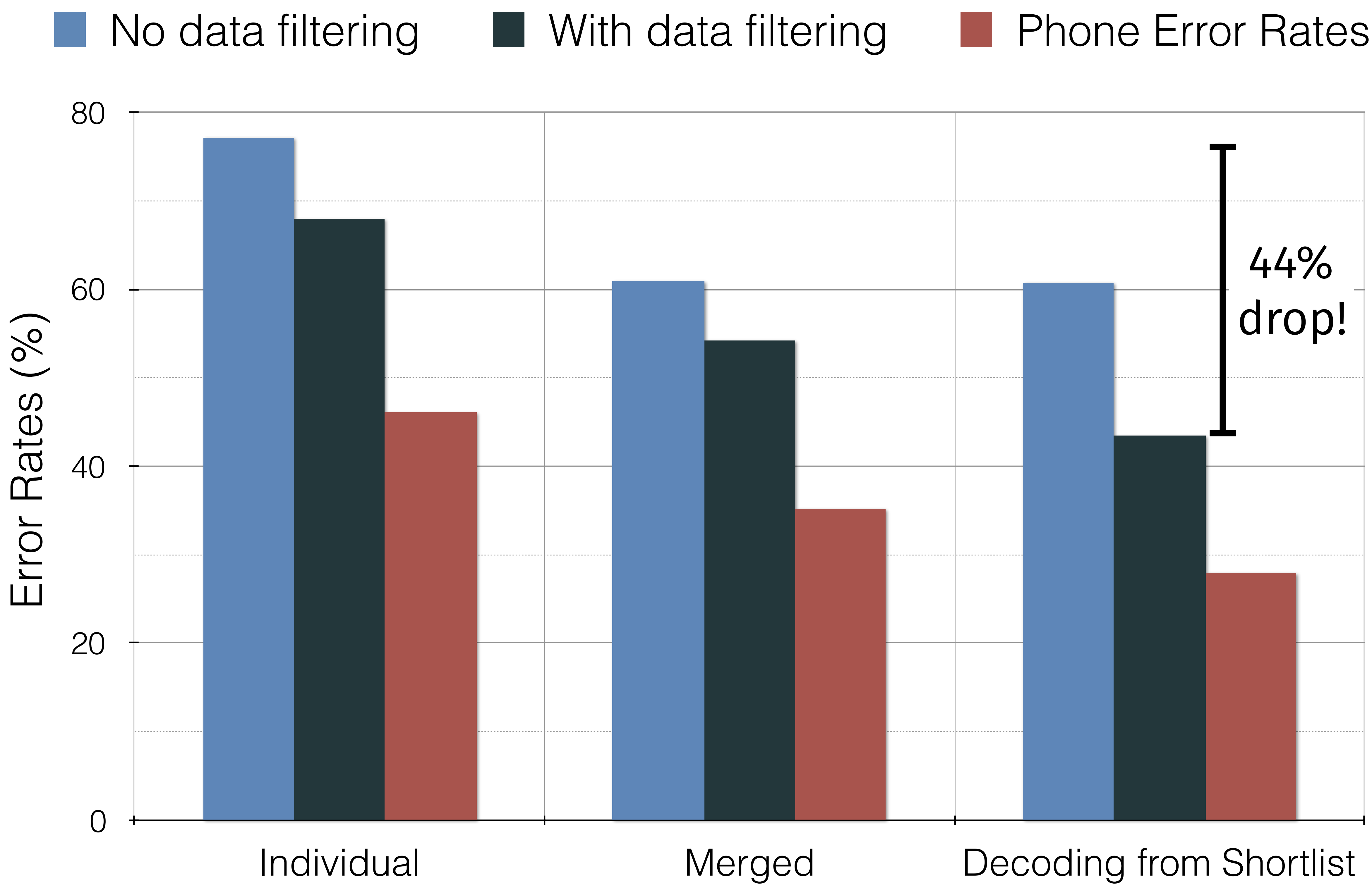
Strike a pose some pizza

Probabilistic phone-based transcriptions derived from alignments[1]

---

[1][P. Jyothi & Hasegawa-Johnson, *Interspeech*-15]

Transcription Error Rates



[Jyothi & Hasegawa-Johnson *Interspeech*-15]

Adapting ASR Systems using Mismatched Transcriptions

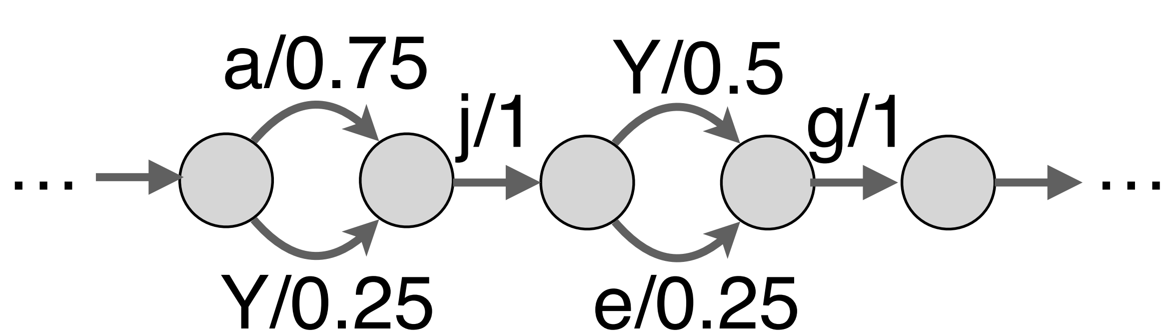
February 7, 2017

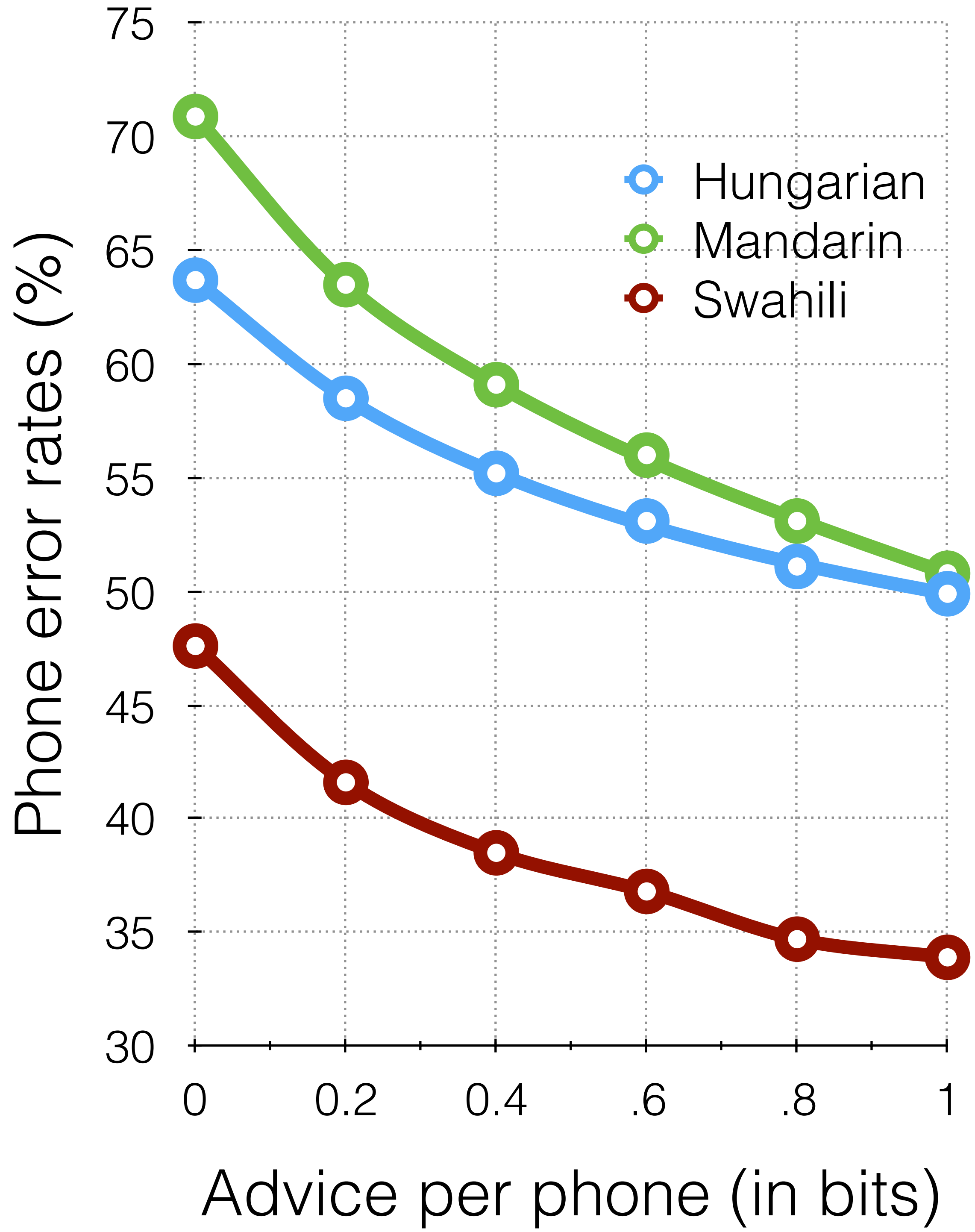
Next Step?

- Respectable accuracy from mismatched transcriptions
  - But can this be leveraged for building ASR systems?
- Plan: Baseline ASR trained on *other* languages will be *adapted* using mismatched transcriptions
  - Baseline could use data-hungry technology like Deep Neural Networks (DNNs)
- Project at 2015 Jelenik Summer Workshop [JSALT '15]
  - Several languages considered: Hungarian, Mandarin, Swahili etc.

More than meets the eye

- Mismatched transcripts too noisy to be used in the traditional way for ASR training
  - Use as *probabilistic* transcripts
  - Measuring additional information in probabilistic transcripts:
    - How error rates fall when more "advice" is made available to the decoder

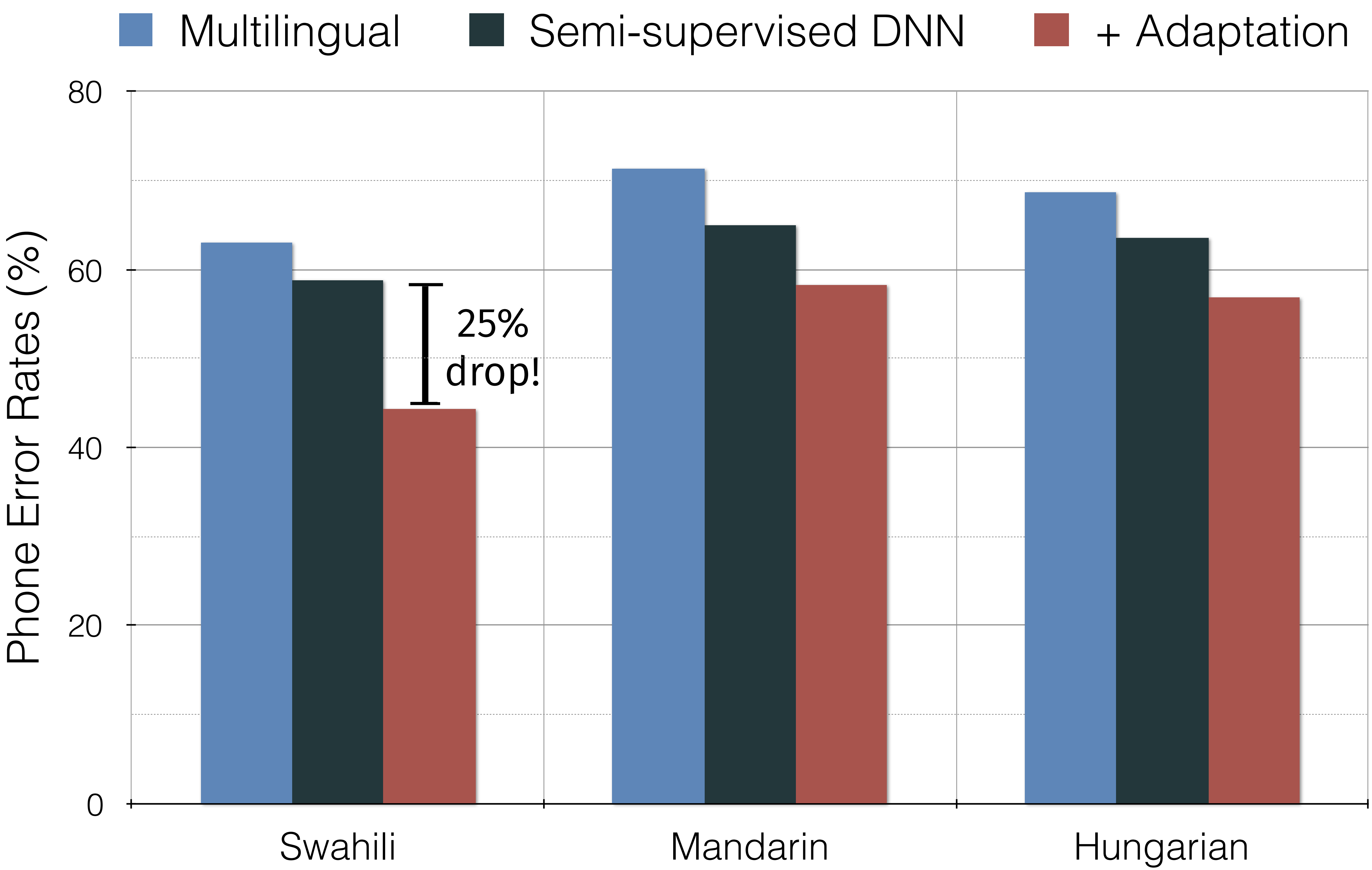




ASR Systems for Comparison

- **Multilingual**: Train on 6 languages (Arabic, Cantonese, Dutch, Hungarian, Mandarin, Urdu) and test on a new target language (Swahili).
- **Semi-supervised DNN:** Transcribe unlabeled audio from the target language using a DNN-based multilingual ASR system and use it to further re-train the DNN models.

Mismatched Transcriptions for ASR



[Liu*, Jyothi* et al. ICASSP' 16]

Native Language Backgrounds of Mismatched Crowds

February 7, 2017

Channel Selection

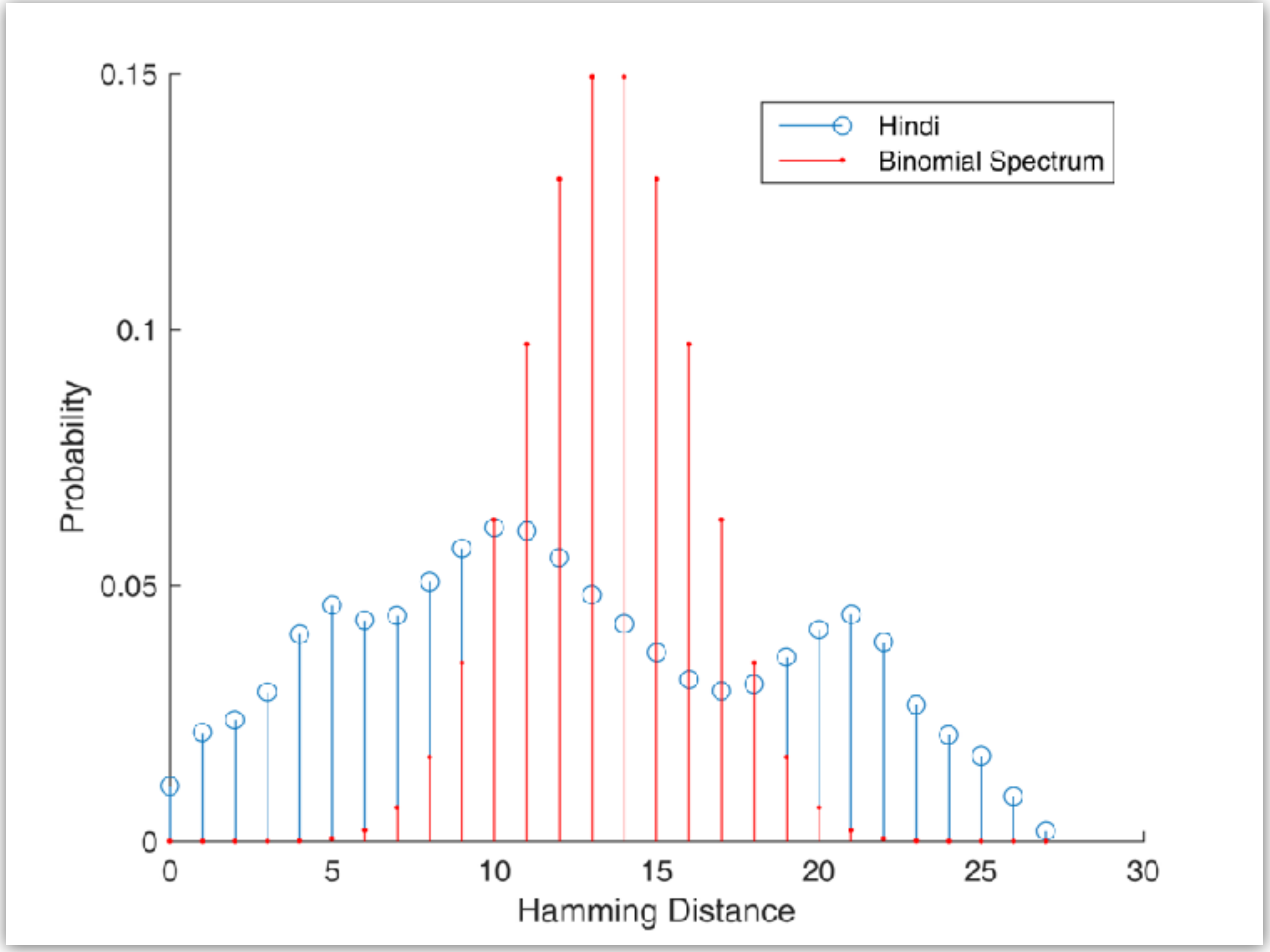
- Can we do better by selecting the language background of the transcribers?
- How should we select the transcribers?
    - Understanding when phones get misperceived
    - How is this correlated with transcribers’ language background

Understanding the Mismatched Channel

- When are two phones confused with each other?
  - If they are "phonologically close" to each other
- Phonological distance between two phones
  - Use ***distinctive features*** (DF) from linguistic theory [Chomsky & Halle, '68] [Phoible '15]
  - Phones as "code words" in the DF-space. Hamming distance measures phone contrast.

37 DFs
nasal
tone
sonorant
labial
trill
front
back
:

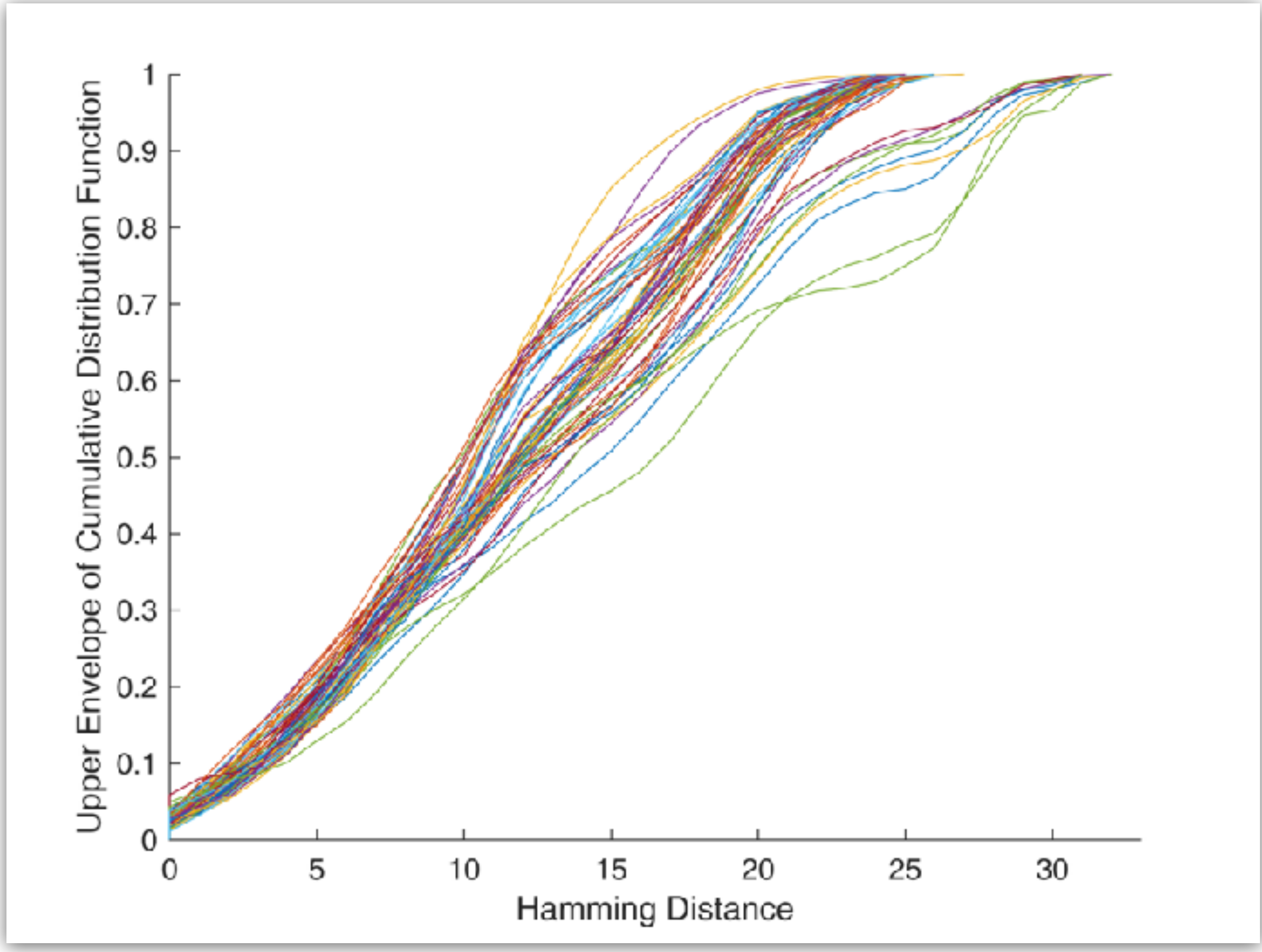
Distance Distribution of the Code



[Varshney, Jyothi & Hasegawa-Johnson *ITA*-16]

Distance Distribution of the Code

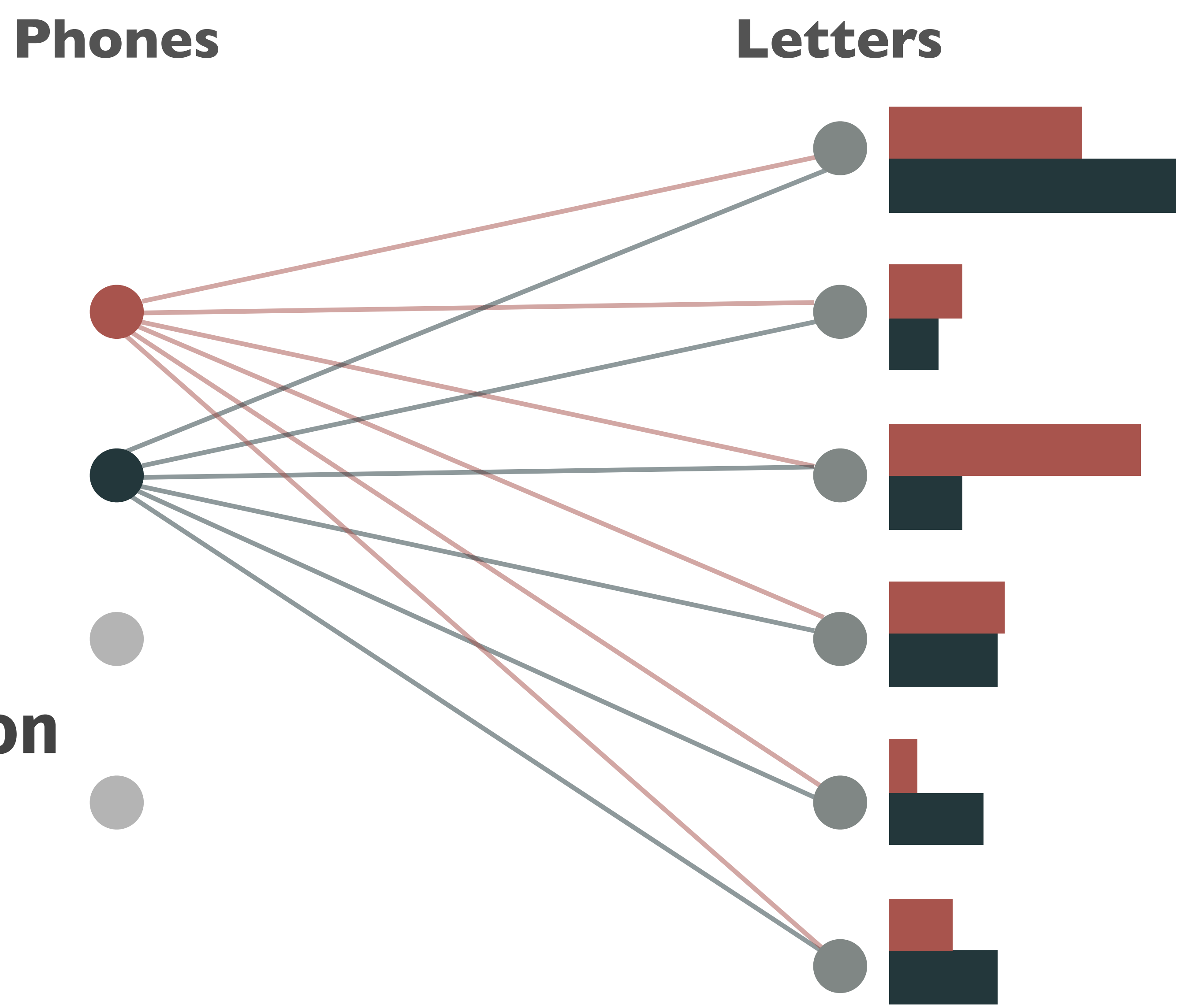
- Codes for different languages exhibit similar distributions



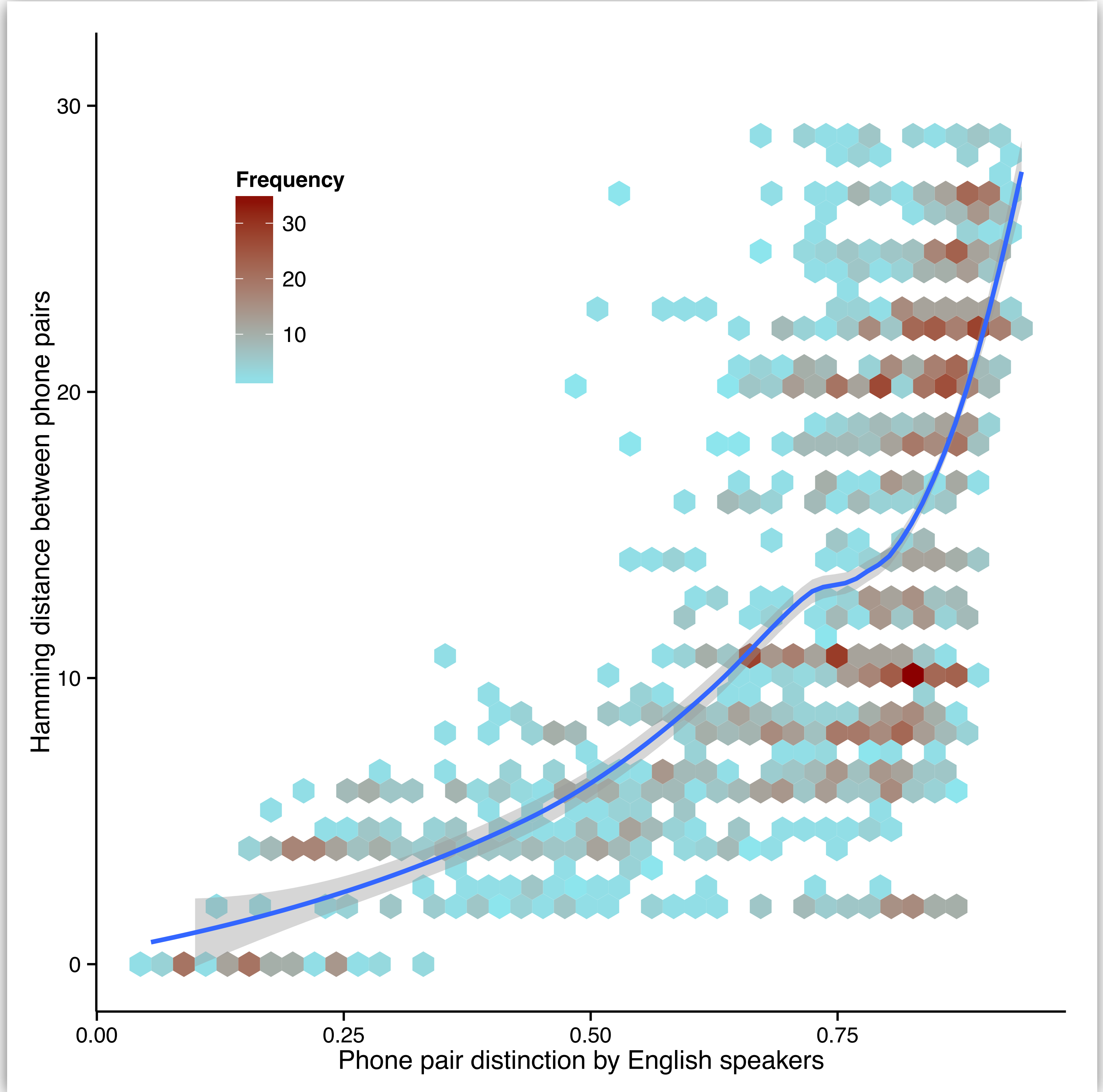
Understanding the Mismatched Channel

- <u>Hypothesis</u>: Phonological distance in the DF-space correlated with *phone confusion* in the mismatched channel

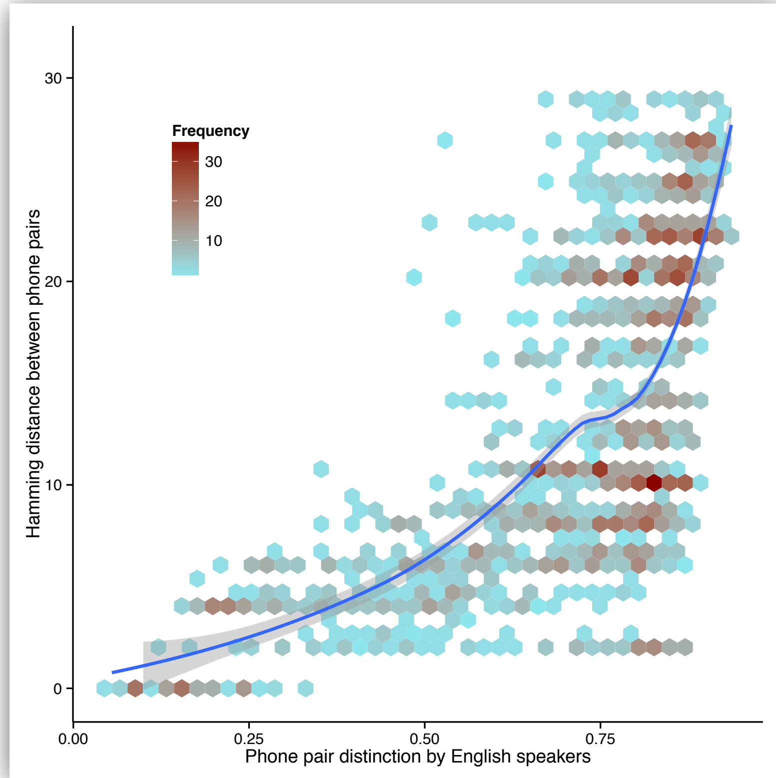
- Phone confusion quantified using the **total variational distance** between the output distributions of the *channel*
  - We call it **phone-pair distinction**

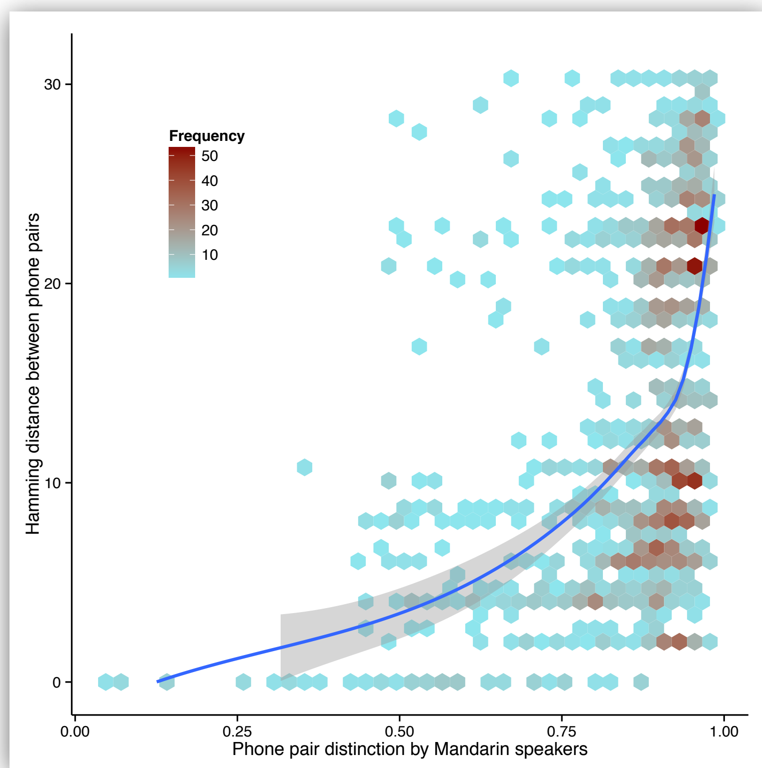


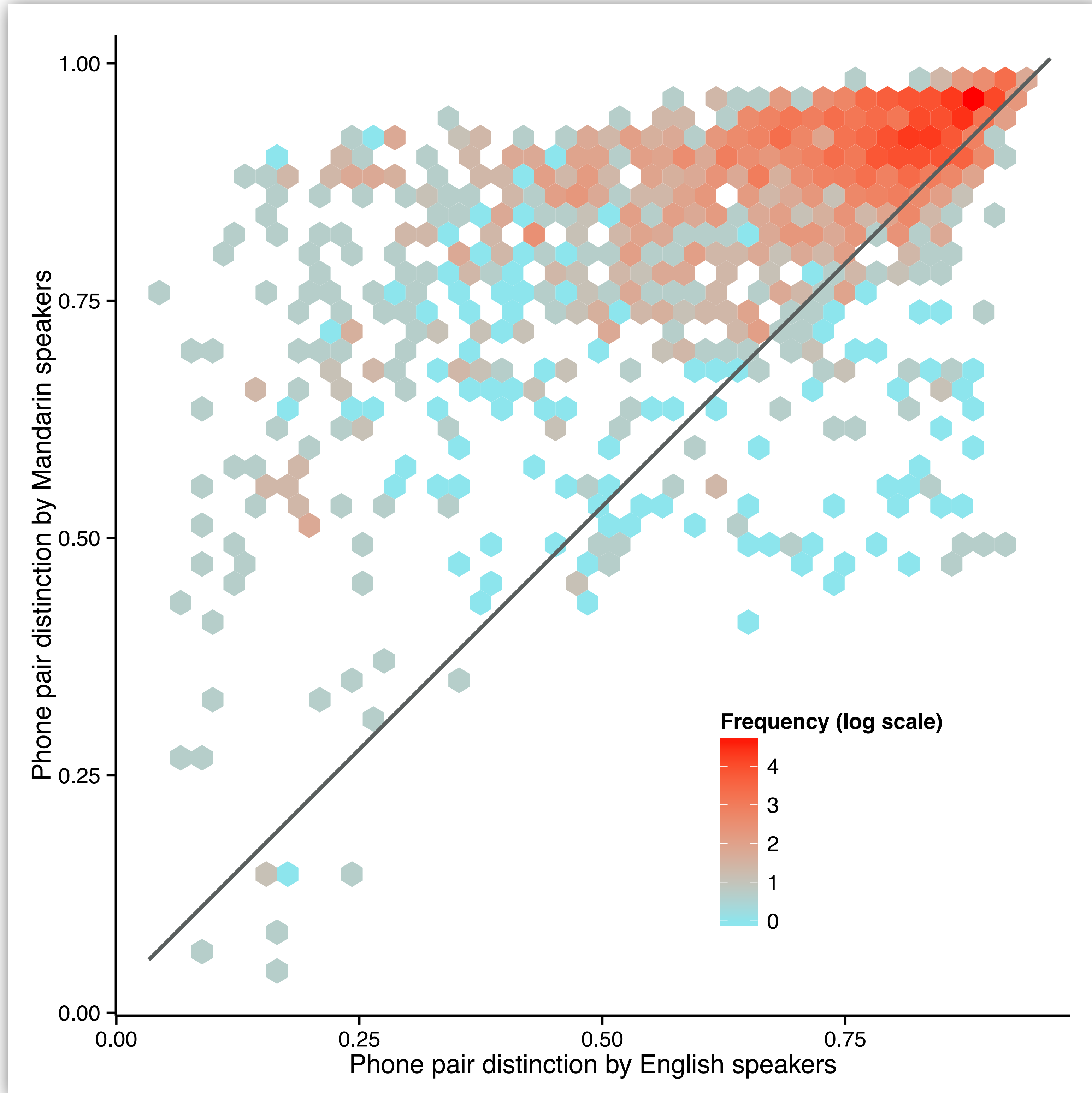
Phone Pair Distinction vs. DF-Distance



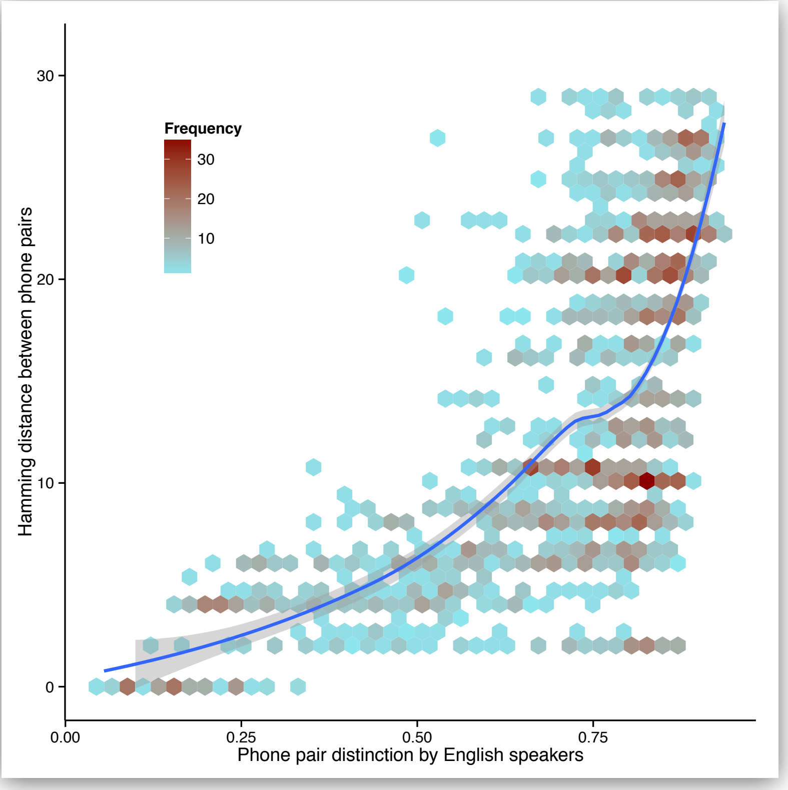
Phone Pair Distinction

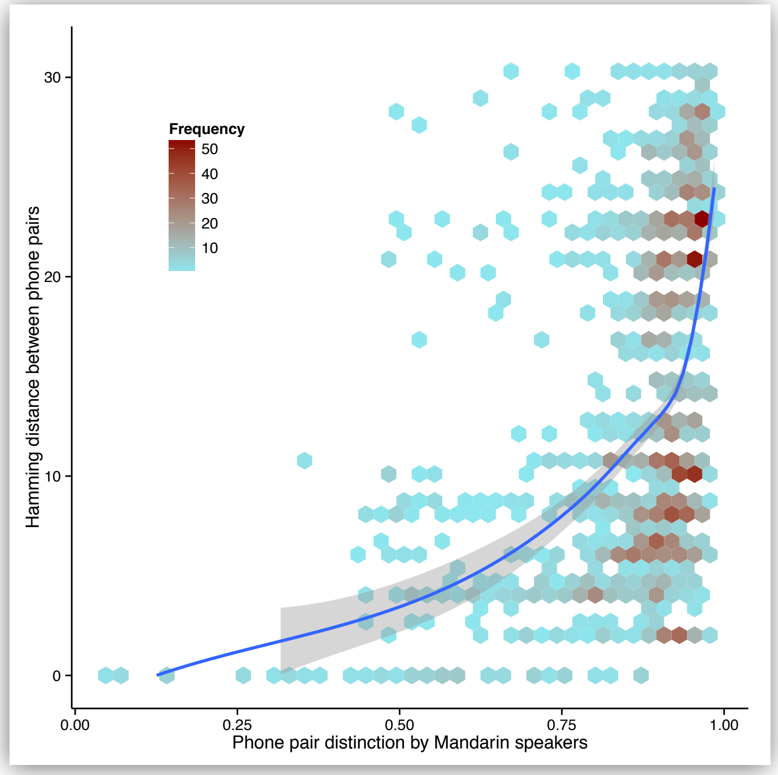






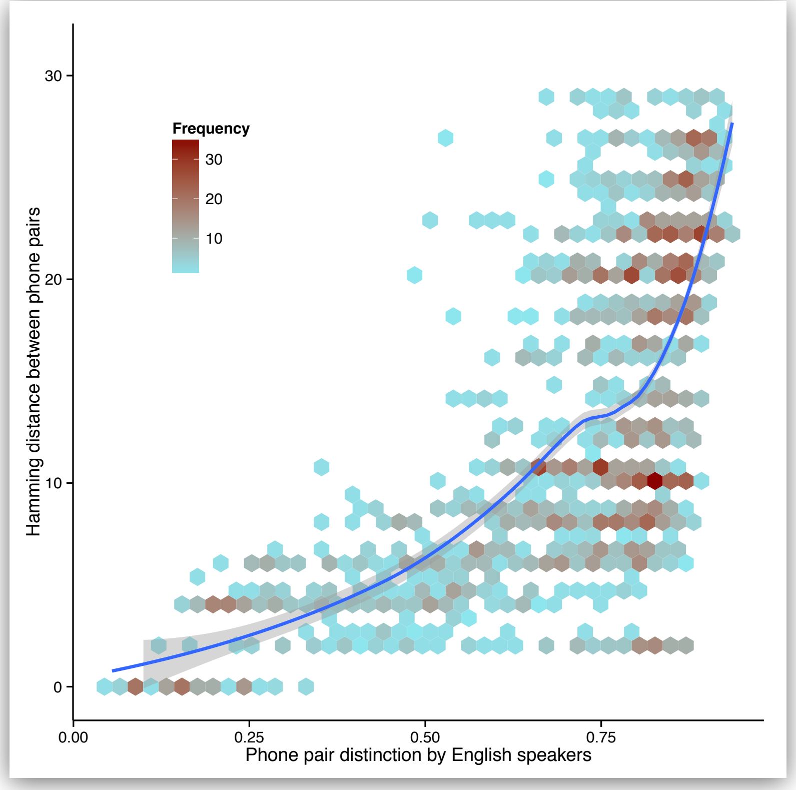
Phone Pair Distinction

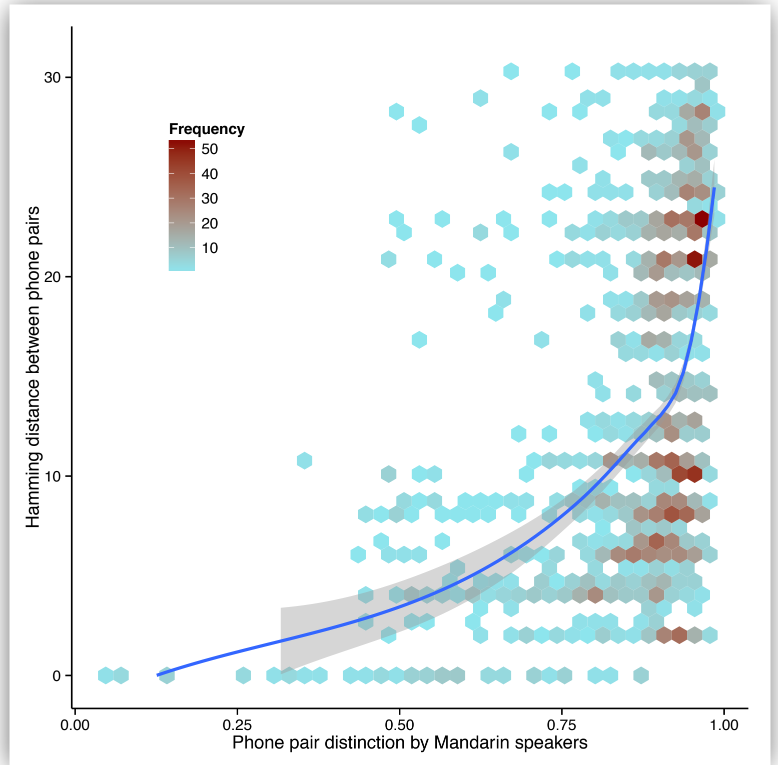




- Clearly, DF-distance positively correlated with phone-pair distinction
- Difference across native language backgrounds
  - Different DFs are prominent in different languages

Phone Pair Distinction



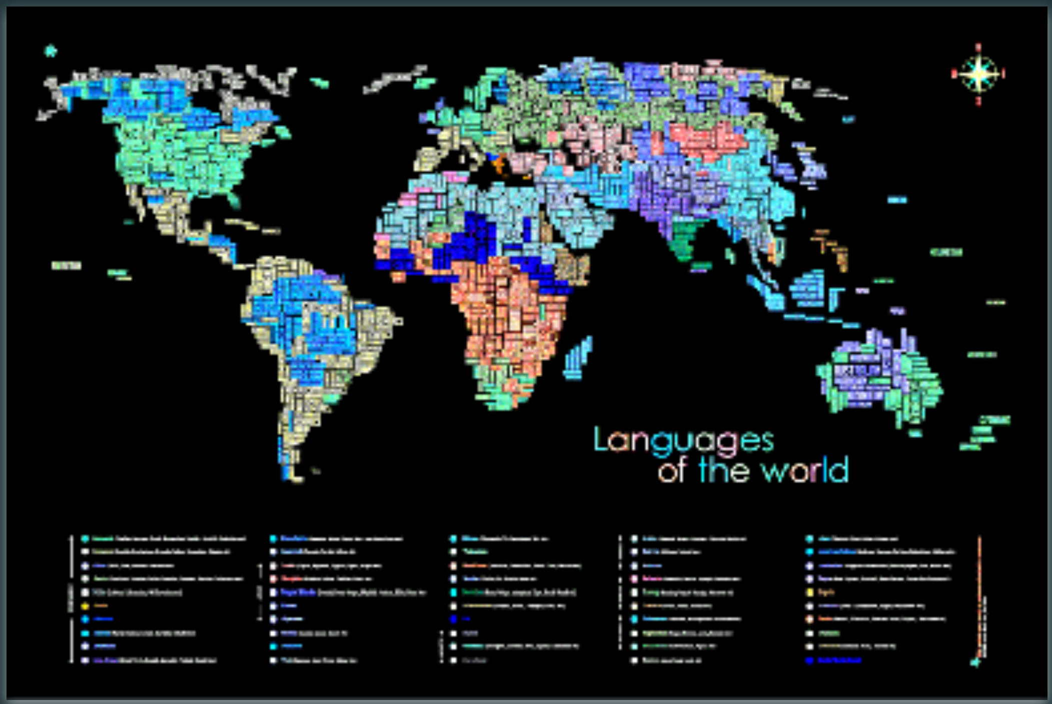


- Clearly, DF-distance positively correlated with phone-pair distinction
- Difference across native language backgrounds
  - Different DFs are prominent in different languages
  - Ongoing work: A model that takes into account DF presence/prominence

Summary

- ASR for low-resource languages presents challenging research problems

- In this talk:
  - Establish the possibility of acquiring speech transcriptions using mismatched crowds
  - Demonstrate the impact of mismatched transcriptions on ASR performance



  - Investigate relation of transcriber native languages with phone confusion

- Future research: Optimally select mismatched transcribers to further improve impact of mismatched transcriptions

Summary

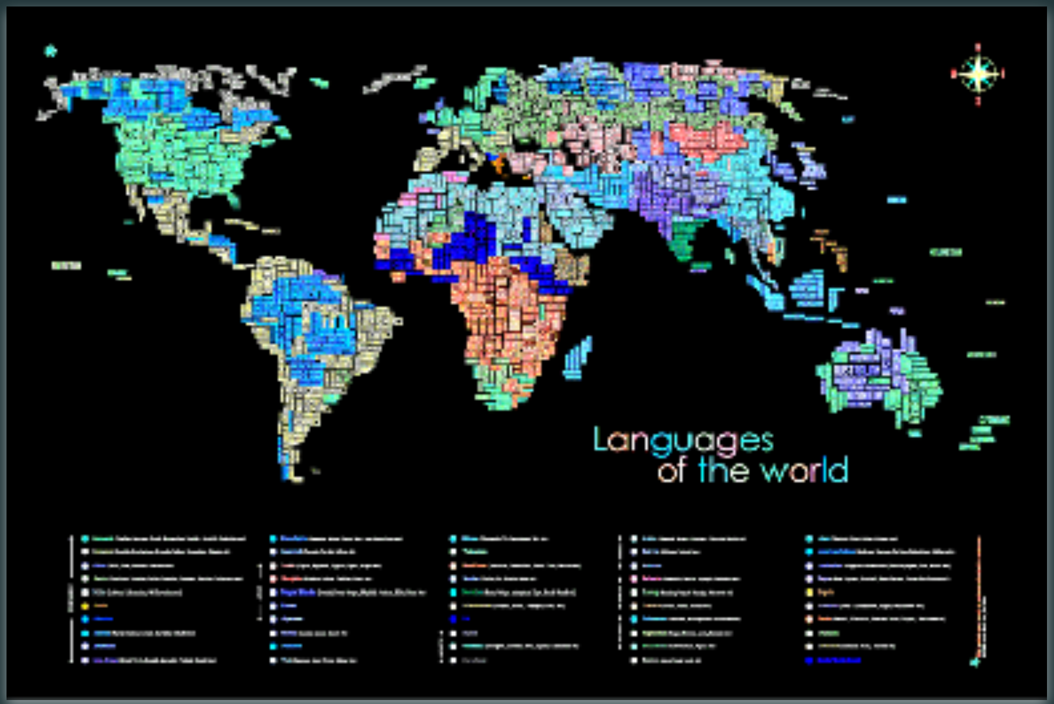
- ASR for low-resource languages presents challenging research problems

- In this talk:

  - Establish the possibility of acquiring speech transcriptions using mismatched crowds

  - Demonstrate the impact of mismatched transcriptions on ASR performance

  - Investigate relation of transcriber native languages with phone confusion



[1]Based on joint works with Mark Hasegawa-Johnson, Lav Varshney and participants at the 2015 Jelinek Summer Workshop.