Learning The Lexicon

A Pronunciation Mixture Model

Ian McGraw

(imcgraw@mit.edu)

Ibrahim Badr

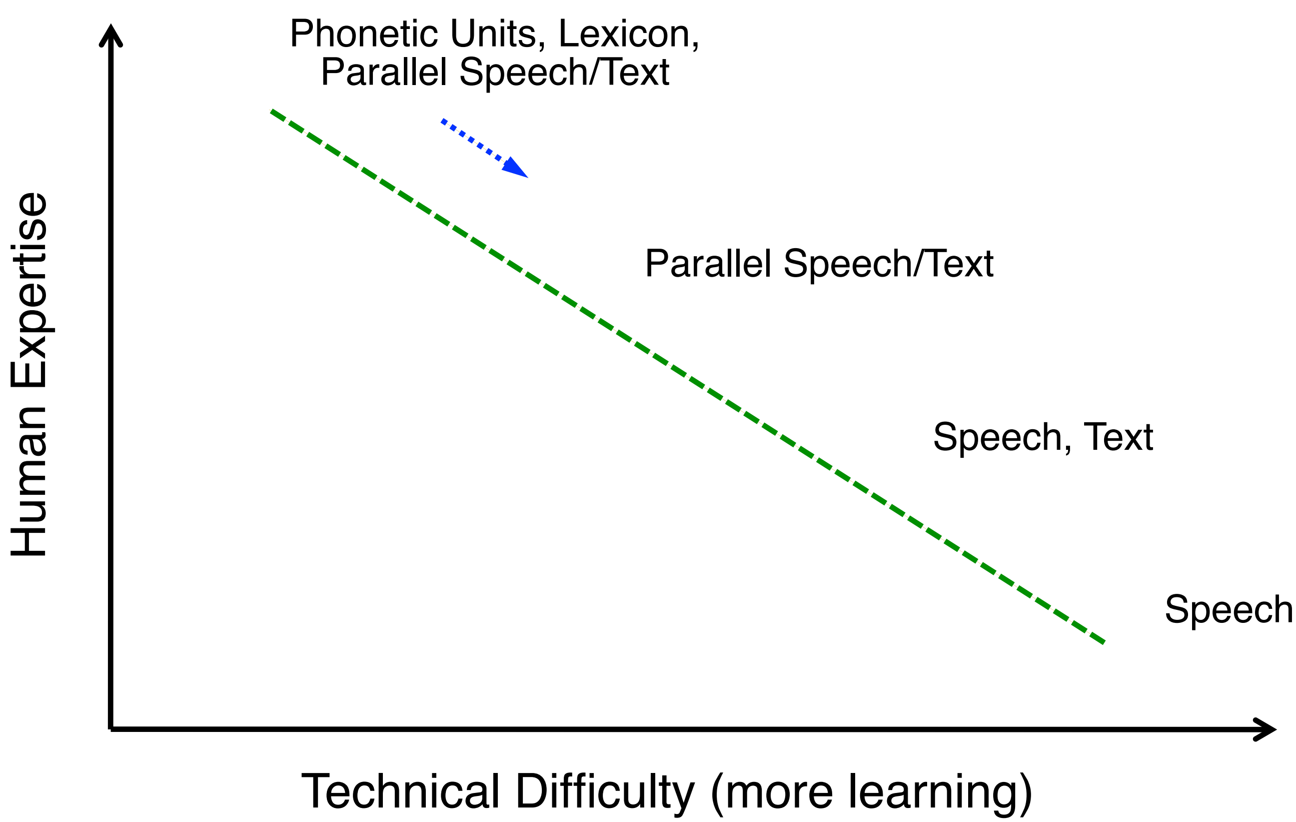
Jim Glass

Computer Science and Artificial Intelligence Lab
Massachusetts Institute of Technology
Cambridge, MA, USA

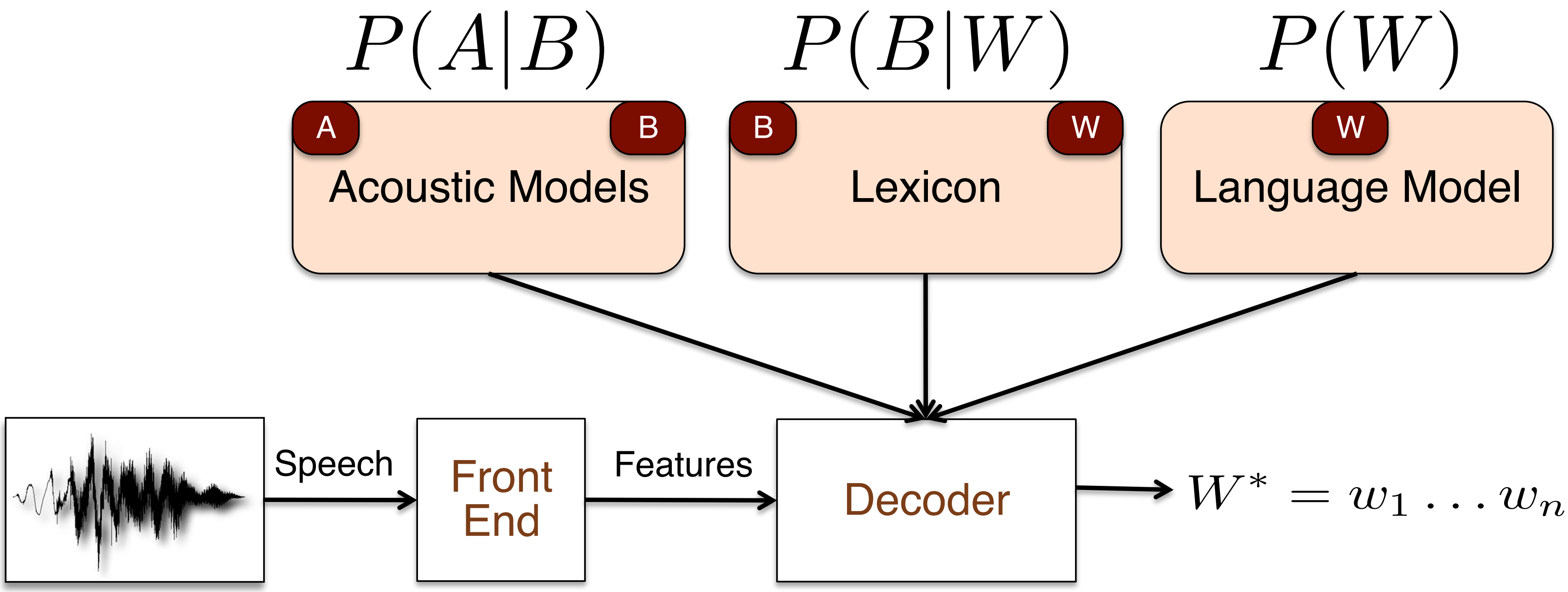
Automatic Speech Recognition

A Perspective on Resources

~98% of the world's languages have not been addressed by resource and expert intensive supervised speech recognition training methods.



Automatic Speech Recognition



Fundamental Equation of ASR?

$$W^* = \underset{W}{\operatorname{argmax}} \max_{B} P(A|B)P(B|W)P(W)$$

Viterbi Approximation!

Stochastic Lexicon

$$P(B|W) = \prod_{j=1} P(\mathbf{b}_j|\mathbf{w}_j)$$

e.g.

$$P(hh\ ax\ w\ ay\ iy|\ \text{hawaii})$$

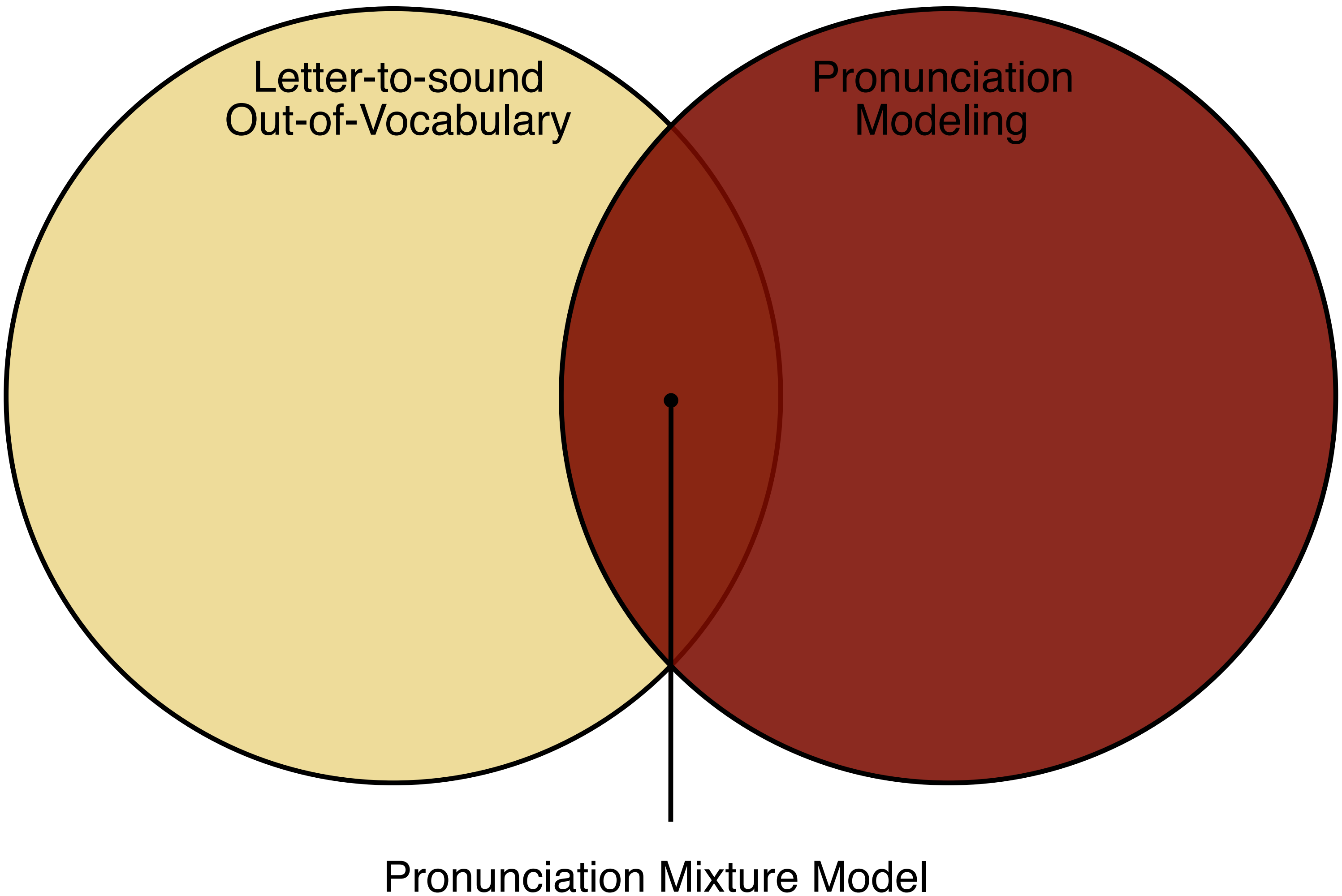
Stochastic Lexicon: A Single Entry

$$P(\mathbf{b}|\mathbf{w})$$

$$1 \qquad \frac{1}{|\{\mathbf{b}_{\mathbf{w}}\}|} \qquad \theta_{\mathbf{b}|\mathbf{w}}$$

Learned Parameters

Pronunciation Related ASR Research



Modeling Pronunciation Variation

Model	Domain	Impact on WER%
Rule learning from manual transcriptions [Riley et al. 1999]	Broadcast news	12.7 ➔ 10.0
	Conversational	44.7 ➔ 43.8
Decision trees + dynamic lexicon [Fosler-Lussier 1999]	Broadcast news	21.4 ➔ 20.4
Knowledge-based rules + FST weight learning [Hazen et al. 2002]	Weather queries	12.1 ➔ 11.0

Phonological Rules

A set of generic rewrite rules convert a baseform from phonemes to its possible variants at the phone level.

Example:

nearest tornado

↓

Phoneme: n ih r ax s td # t ao r n ey df ow

↓

Phone: n ih r ax s # tcl t er n ey dx ow

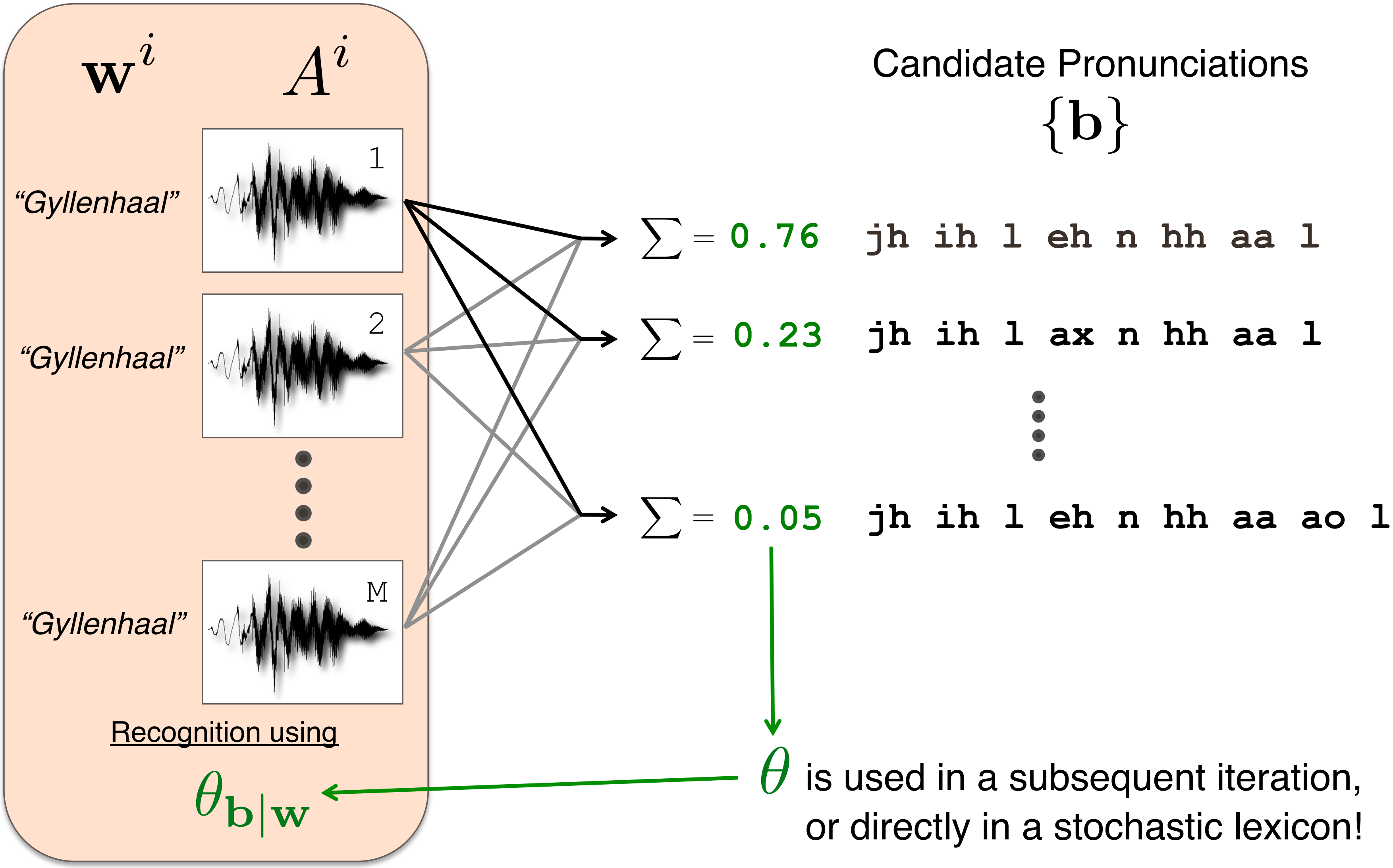
n ih r ax s tcl t # tcl t ao r n ey dcl d ow

…

Expanding the lexicon

Pronunciation Mixture Model

CSAIL



Pronunciation Mixture Model

E-step:

$\overline{\mathrm{M}}_\theta[\mathbf{w}, \mathbf{p}]$ is expected number of times that pronunciation $\mathbf{p}$ is used for word $\mathbf{w}$.

M-step:

$\theta^*_{\mathbf{p}|\mathbf{w}}$ is the normalized $\overline{\mathrm{M}}_\theta[\mathbf{w}, \mathbf{p}]$ across all pronunciations for a given word $\mathbf{w}$.

Continuous Speech PMM Example

	what	do	you	know
0.5584	w ah tcl	dcl d uw	y uw	n ow
0.2250	w ah tcl	dcl d	y uw	n ow
0.0645	w ah tcl	dcl d uw	y ax	n ow
0.0434	w ah tcl t	dcl d uw	y uw	n ow
0.0376	w ah tcl	dcl d ow	y uw	n ow
0.0244	w ah tcl	dcl d ax	y uw	n ow
0.0219	w ah tcl	dcl d iy	y uw	n ow
0.0097	w ah tcl	dcl d	y ax	n ow
0.0083	w ah tcl t	dcl d uw	y ax	n ow
0.0063	w ah tcl	dcl jh	y uw	n ow

	thank	you
0.4954	th ae ng kcl k	y ax
0.4891	th ae ng kcl k	y uw
0.0068	th ae ng kcl k	y ah
0.0035	th ae ng kcl k	y axr
0.0010	th ae ng kcl	y ax
0.0010	th ae ng kcl	y uw
0.0010	th ae ng kcl k	y ow
0.0007	th ae ng kcl k	y el
0.0004	th ae ng kcl k	y aa uw
0.0004	th ae ng kcl k	y aw

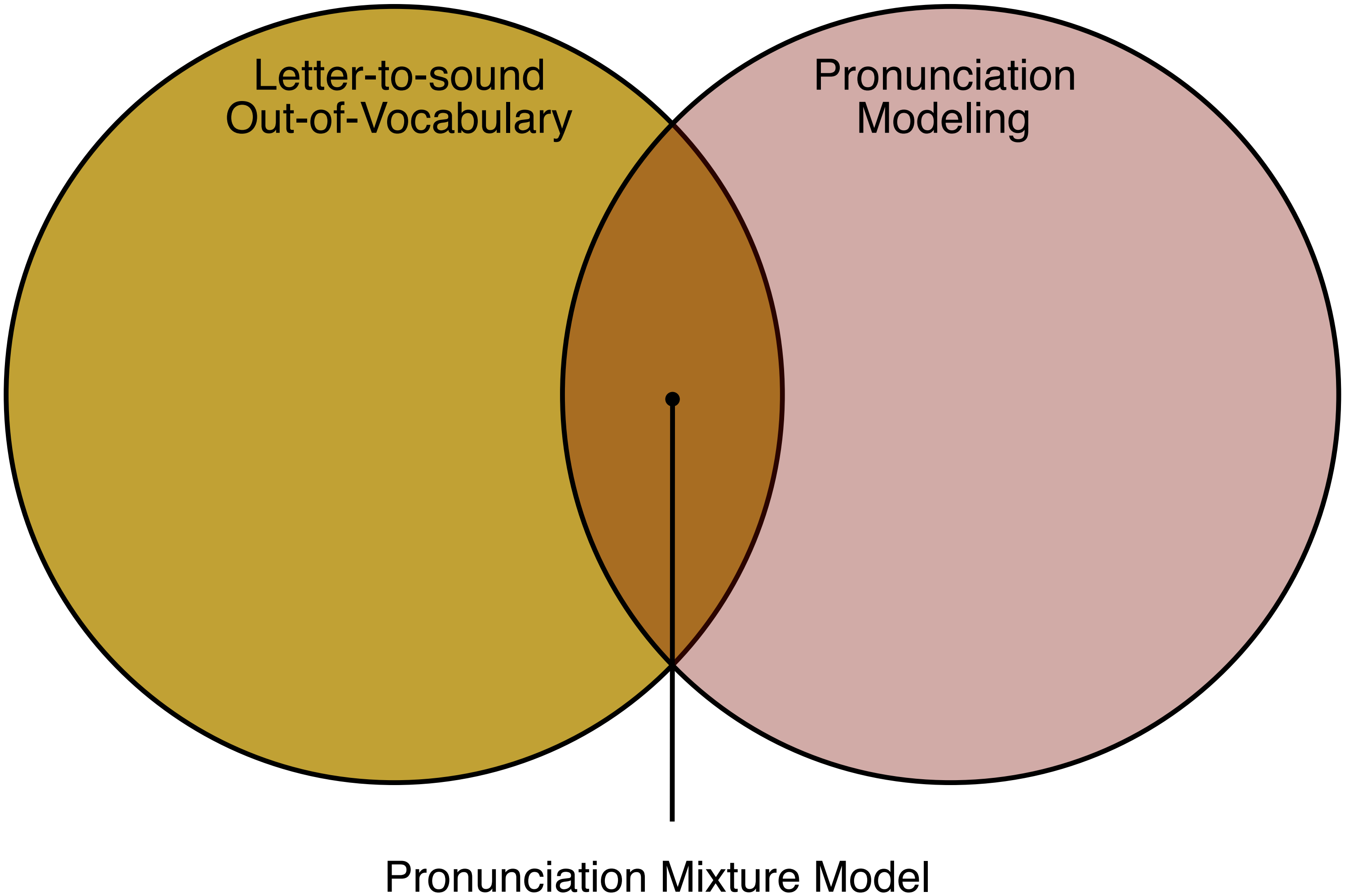
$$\theta_{y\ uw|\text{you}}$$

$$\theta_{y\ ax|\text{you}}$$

$$\theta_{y\ ah|\text{you}}$$

Normalize!

Pronunciation Related ASR Research



Initializing a PMM: Choosing the Support

- **Expert Pronunciations**
  - The SLS dictionary is based on PronLex
    - **Contains around 150,000 words**
    - **Has an average of 1.2 pronunciations per word.**
    - **Supplementary phonological rules expand pronunciations.**
    - **Expanded lexicon has an average of 4.0 pronunciations per word.**
- **Letter-to-sound L2S System**
  - Joint sequence models [Bisani and Ney, 2008]
    - **Graphonemes: Train on expert lexicon.**
    - **Graphones: Train on expanded expert lexicon.**

Joint-sequence Models for L2S

CSAIL

$$\mathbf{b}^* = \operatorname*{argmax}_{\mathbf{b}} P(\mathbf{w}, \mathbf{b})$$

	w	=	c	o	u	p		l	e
	b	=	k	ah		p	ax	l	
		=	k		ah	p	ax	l	
S(w,b)	g_1	=	c/k	o/ah	u/	p/p	/ax	l/l	e/
	g_2	=	c/k	o/	u/ah	p/p	/ax	l/l	e/

$$P(\mathbf{w}, \mathbf{b}) = \sum_{\mathbf{g} \in S(\mathbf{w}, \mathbf{b})} P(\mathbf{g}) \approx \max_{\mathbf{g} \in S(\mathbf{w}, \mathbf{b})} P(\mathbf{g})$$

EM used to infer alignments and build an M-gram over multigrams.

Non-singular Graphones

Parameters:

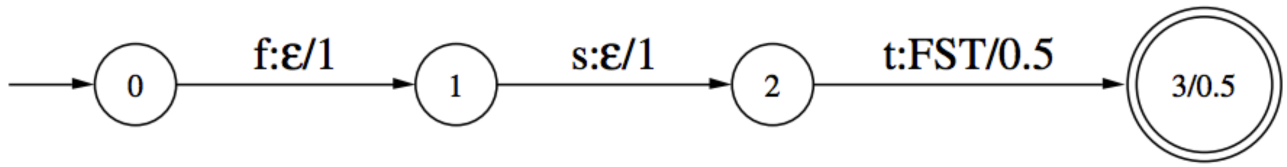
L = the number of graphemes

R = the number of phonetic units

M = language model context size

w	=	c	o	u	p		l	e
b	=	kcl_k	ah		pcl_p	ax	l	
	=	kcl_k		ah	pcl_p	ax	l	
g_1	=	c/kcl_k	o/ah	u/	p/pcl_p	/ax	l/l	e/
g_2	=	c/kcl_k	o/	u/ah	p/pcl_p	/ax	l/l	e/

Weighted Finite State Transducers



$$J \leftrightharpoons P(\mathbf{w}, \mathbf{b})$$

$$J_{\hat{\mathbf{w}}} \leftrightharpoons P(\hat{\mathbf{w}}, \mathbf{b})$$

$G \leftrightharpoons$ Language Model

$L \leftrightharpoons$ Stochastic Lexicon

$P \leftrightharpoons$ Phonological Rules

$C \leftrightharpoons$ Context Dependent Labels

Standard Recognition

$$R = C \circ P \circ L \circ G$$

PMM Training

$$\mathrm{S}_{W^i} = J_{\mathbf{w^i_1}} \# J_{\mathbf{w^i_2}} \# \ldots \# J_{\mathbf{w^i_{|W|}}}$$

$$R^i = C \circ P \circ S_{W^i}$$

$$R^i = C \circ S_{W^i}$$

Recognition Output During Training

(Effectively)

	#	what	#	do	#	you	#	know	#
0.5584	#	w ah tcl	#	dcl d uw	#	y uw	#	n ow	#
0.2250	#	w ah tcl	#	dcl d	#	y uw	#	n ow	#
0.0645	#	w ah tcl	#	dcl d uw	#	y ax	#	n ow	#
0.0434	#	w ah tcl t	#	dcl d uw	#	y uw	#	n ow	#
0.0376	#	w ah tcl	#	dcl d ow	#	y uw	#	n ow	#
0.0244	#	w ah tcl	#	dcl d ax	#	y uw	#	n ow	#
0.0219	#	w ah tcl	#	dcl d iy	#	y uw	#	n ow	#
0.0097	#	w ah tcl	#	dcl d	#	y ax	#	n ow	#
0.0083	#	w ah tcl t	#	dcl d uw	#	y ax	#	n ow	#
0.0063	#	w ah tcl	#	dcl jh	#	y uw	#	n ow	#

	#	thank	#	you	#
0.4954	#	th ae ng kcl k	#	y ax	#
0.4891	#	th ae ng kcl k	#	y uw	#
0.0068	#	th ae ng kcl k	#	y axr	#
0.0035	#	th ae ng kcl k	#	y el	#
0.0010	#	th ae ng kcl	#	y ax	#
0.0010	#	th ae ng kcl	#	y uw	#
0.0010	#	th ae ng kcl k	#	y ow	#
0.0007	#	th ae ng kcl k	#	y ah	#
0.0004	#	th ae ng kcl k	#	y aa uw	#
0.0004	#	th ae ng kcl k	#	y aw	#

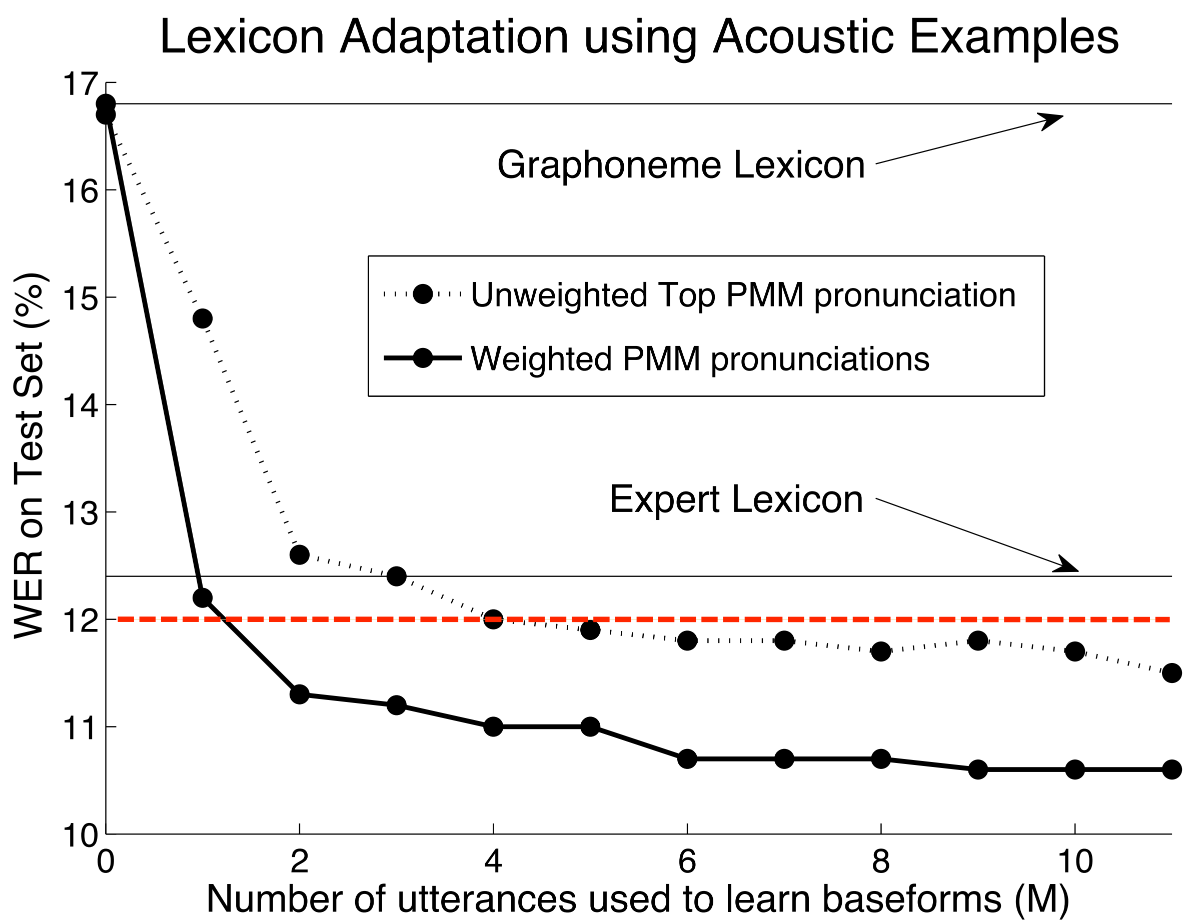
SUMMIT Recognizer Setup

- **Landmark-based speech recognizer**
  - MFCC averages are taken at varying durations around hypothesized boundaries
  - 112-dimensional feature vectors are whitened with a PCA rotation
  - 50 principal components are kept
- **Context dependent acoustic models**
  - Up to 75 diagonal gaussian mixture components each
  - Maximum-likelihood back-off models are trained on a corpus of telephone speech
- **Lexicon**
  - Expert dictionary based on PronLex
  - ~150,000 words for training graphones/graphonemes
  - All lexicons limited to training set words for testing

Isolated Words: Experimental Setup

- **Phonebook Corpus**
  - Isolated words spoken over the telephone by Americans
  - 2,000 words chosen randomly from the set that had at leas 13 speakers
  - For each word, 1 male and 1 female utterance was held out for testing
  - The remaining 22,000 utterances were used for training
- **Lexicon**
  - The 2,000 test words were removed from the lexicon
  - A simple edit distance criterion was used to prune similar words
  - A 5-gram graphoneme language model was trained
- **Testing**
  - We test lexicons trained using varying amounts of acoustic data
  - We compare with two baselines:
    - **Expert Baseline WER: 12.4%**
    - **Graphoneme L2S WER: 16.7%**

Isolated Word: Experimental Results



amazonmechanical turk
Artificial Artificial Intelligence

Collect 10 pronunciations for each word.

Train a PMM just on these pronunciations.

Analysis

83% of the top pronunciations are identical

Word	Dictionary Baseform	Top PMM Pronunciation
parishoners [sic]	p **AE** r ih sh ax n er z	p **AX** r ih sh ax n er z
traumatic	tr r **AO** m ae tf ax kd	tr r **AX** m ae tf ax kd
winnifred	w ih n ax f r **AX** dd	w ih n ax f r **EH** dd
crosby	k r ao **Z** b iy	k r aa **S** b iy
melrose	m eh l r ow **Z**	m eh l r ow **S**
arenas	**ER** iy n ax z	**AX R** iy n ax z
billowy	b ih l **OW** iy	b ih l **AX W** iy
whitener	w ay **TF AX** n er	w ay **TD** n er
airsickness	eh r **SH** ih kd n **EH** s	eh r **S** ih kd n **AX** s
Isabel	**AX S AA** b eh l	**IH Z AX** b eh l

Continuous Speech: Experimental Setup

- **Jupiter: Weather Query Corpus**
  - Short queries (average of ~6 words in length)
  - We prune utterances with non-speech artifacts out of the corpus
  - We use a training set containing 76.68 hours of speech
  - We use a test set containing 3.18 hours and a dev set of .84 hours
  - The acoustic models are well matched to the training set
- **Lexicon & LM**
  - The lexicons in these experiments contain only training set words
  - A trigram was trained on the training set transcripts
- **Testing**
  - We provide a baseline based only on the letter-to-sound pronunciations
  - We decode using the expert dictionary to give us a baseline
  - We weight the expert lexicon using the PMM training
  - We train a 5-gram graphone and graphoneme-based PMM

Continuous Speech: Experimental Results

	WER
Graphoneme L2S	11.2
Expert	9.5
Expert PMM	9.2
Phoneme PMM	8.3
Phone PMM	8.2

Analysis

Word	Dictionary Baseform	Top PMM Pronunciation
already	ao **L** r eh df iy	aa r eh df iy
antarctica	ae nt aa r **KD** t ax k ax	ae nt aa r tf ax k ax
asked	ae s **KD** t	ae s td
barbara	b aa r b **AX** r ah	b aa r b r ax
bratislava	b r **AA** tf ax s l aa v ax	b r tf ax z l aa v ax
clothes	k l ow **DH** z	k l ow z

Lexicon Sizes

Weather Lexicon	Avg # Prons	# States	# Arcs	Size
Expert	1.2	32K	152K	3.5 MB
Phoneme PMM	3.15	51K	350K	7.5 MB
Phone PMM	4.0	47K	231K	5.2 MB

Varying Initialization Parameters

Singular Graphones and variable M					
	M=1	M=2	M=3	M=4	M=5
LM FST Size	28K	64K	624K	3.1M	9.4M
WER Using Graphones Alone (T=.01)					
PPW	3.9	14.2	11.4	7.5	5.9
WER Dev.	71.4	20.5	14.3	13.1	12.5
WER Test	74.9	17.6	11.7	10.9	10.2
WER Using Graphone-based PMM (T=.01)					
PPW	6.08	4.0	3.5	3.1	3.0
WER Dev.	17.7	10.6	10.6	10.5	10.7
WER Test	15.6	8.4	**8.2**	8.1	8.2

Varying Initialization Parameters

CSAIL

Graphones with $M = 5$ and variable L-to-R					
	1-to-1	1-to-2	1-to-3	2-to-2	3-to-3
LM FST Size	9.4M	12M	12M	25M	21M
WER Using Graphones Alone (T=.01)					
PPW	5.9	5.7	5.7	5.4	5.5
WER Dev.	12.5	12.2	12.4	12.2	12.3
WER Test	10.2	10.2	10.3	10.4	10.1
WER Using Graphone-based PMM (T=.01)					
PPW	3.0	3.0	2.9	2.9	2.9
WER Dev.	10.7	10.6	10.5	10.4	10.6
WER Test	**8.2**	8.2	8.2	8.2	8.2

Conclusions

CSAIL

- PMM: Maximum-likelihood lexicon learning
- Flexible initialization from experts or L2S
- Requires no additional resources
- Produces better-than-expert pronunciations

Future Directions

- Apply to other languages
- Train in tandem with acoustic models (no phonological rules)
- Learn the lexicon from scratch?