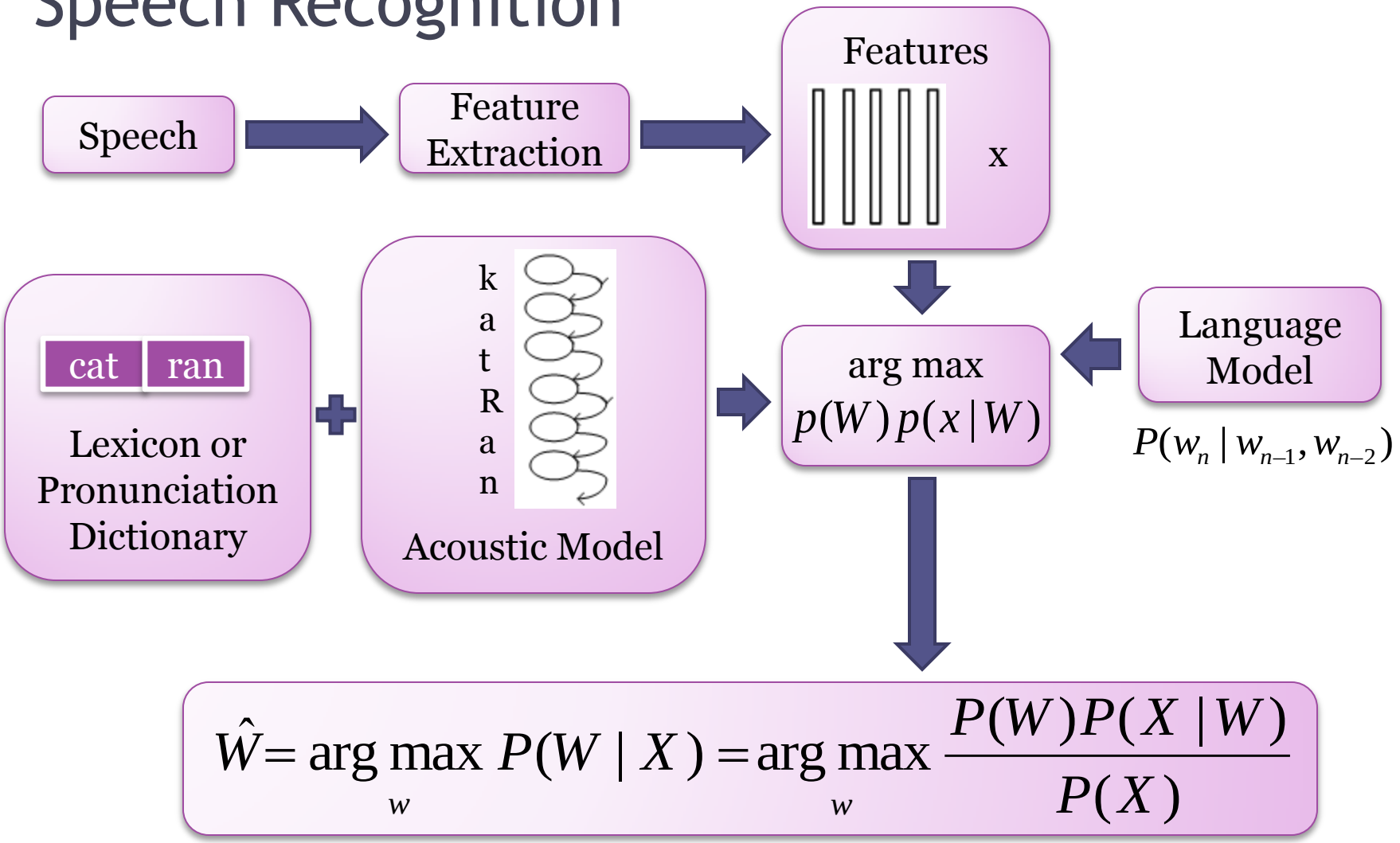
Low Cost Lexicon

Nagendra Kumar Goel, Samuel Thomas, Pinar Akyazi

Speech Recognition



Cost of Developing a New Language

- Transcribed audio data
  - Subspace acoustic models (UBM's) need less data
- Text data for language modeling
  - Obtain from the web if possible
- Pronunciation Lexicon
  - Qualified phoneticians are expensive
  - Phoneticians may make mistakes
  - Conversational (callhome) English has 4.6% OOV rate for a 5K lexicon and 0.4% for a 62K lexicon
  - Try to guess pronunciation given a limited lexicon and audio

Estimating Pronunciations

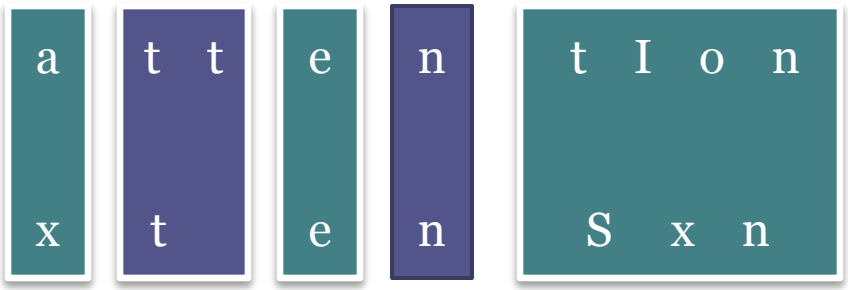
- Ideal Situation will be to just estimate all the pronunciations for the word that maximize the likelihood given the audio

$$\hat{Prn} = \arg\max_{Prn} P(X \mid Prn)$$

- There are words for which spoken audio is not available but they need to exist in the recognizer.
- Multiple pronunciations have not yet significantly inproved the performance
- This objective function needs a lot of regularization

Estimating Pronunciation from Graphemes

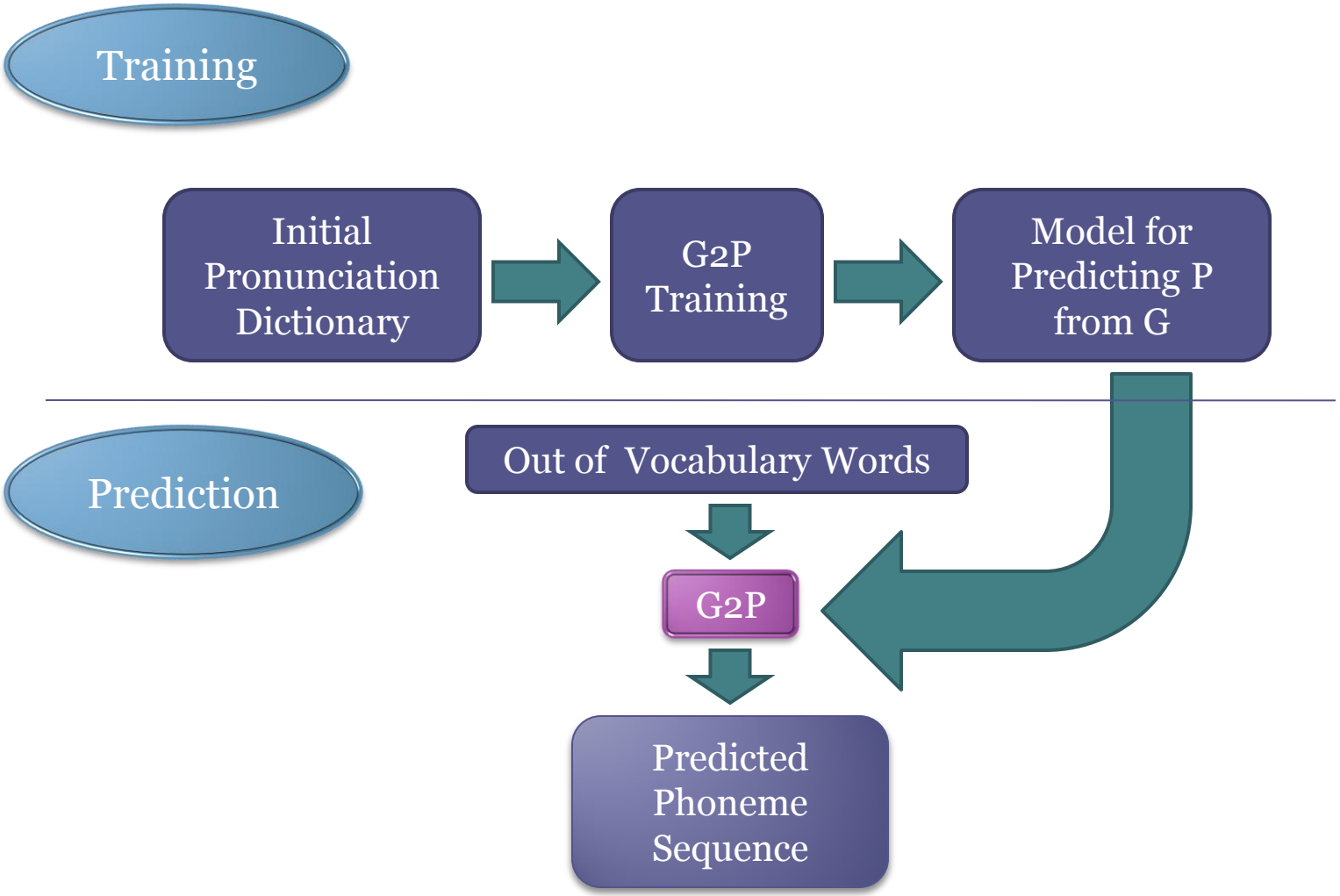
- One way is to guess the pronunciation from the orthography of the word (e.g. Bisani & Ney)
- Iterative process based on grapheme/phoneme alignment
  - Start with an initial set of graphone probabilities.



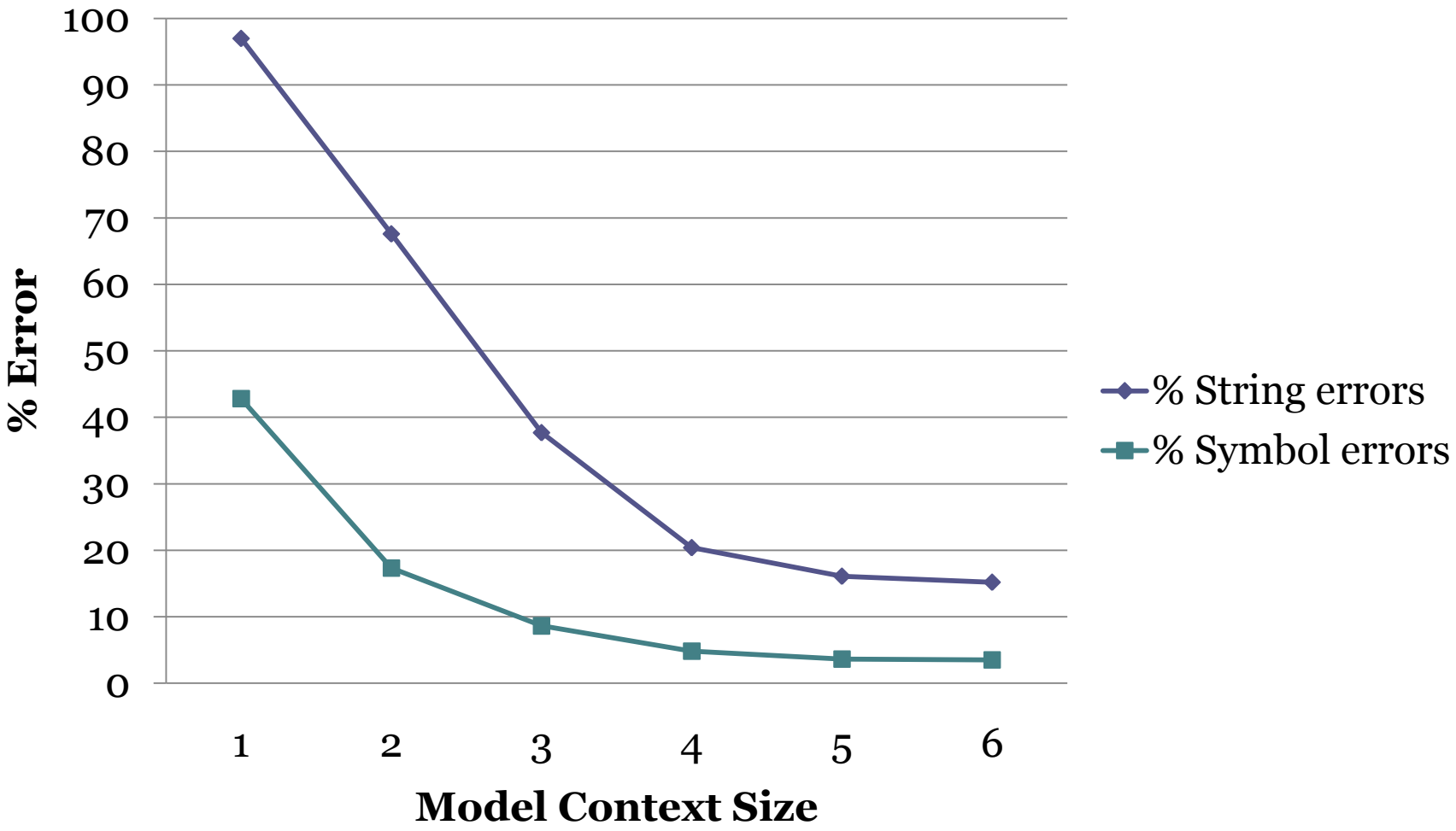
$$\hat{Prn} = \arg\max_{Prn} P(W, Prn)$$

  - Use the probabilities to realign graphones with phones on training data.
  - Re-estimate graphone probabilities from the alignments.

Training a Pronunciation Dictionary



G2P Plot for English



Estimating Pronunciations…

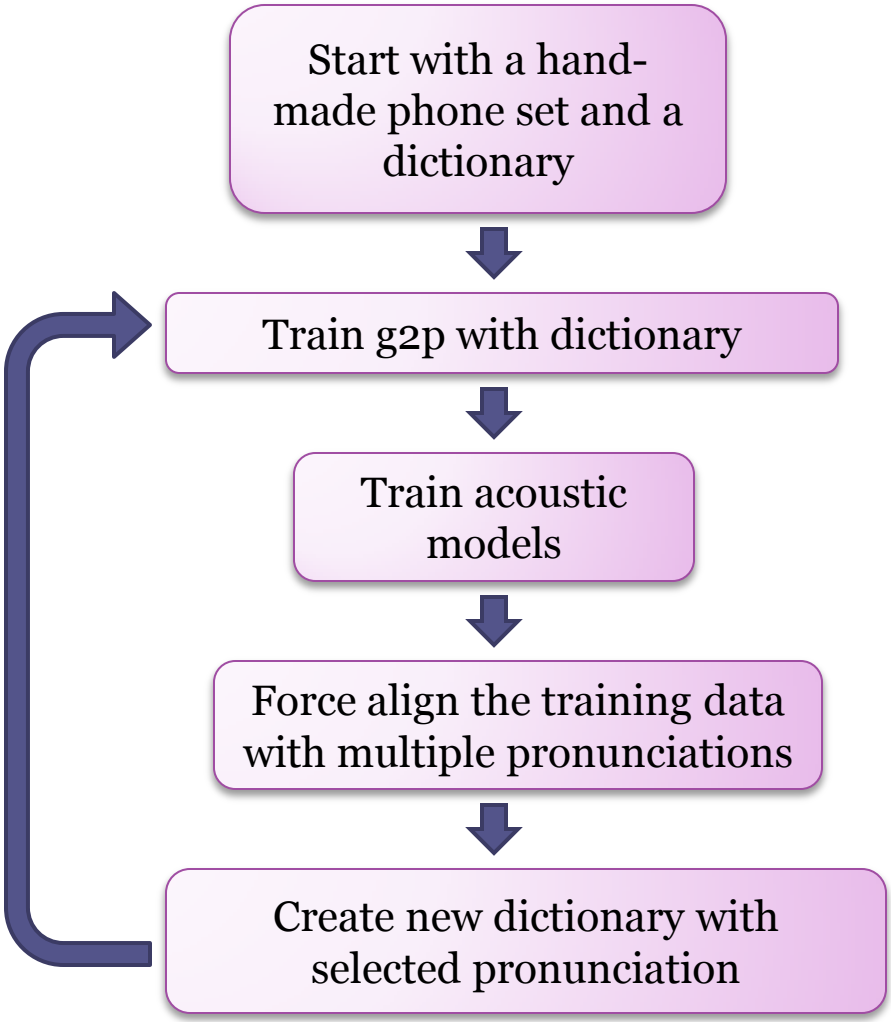
- If the audio recording is also available, that can be used to augment the estimates

$$\hat{Prn} = \arg\max_{Prn} P(X \mid Prn) P(Prn \mid W)$$

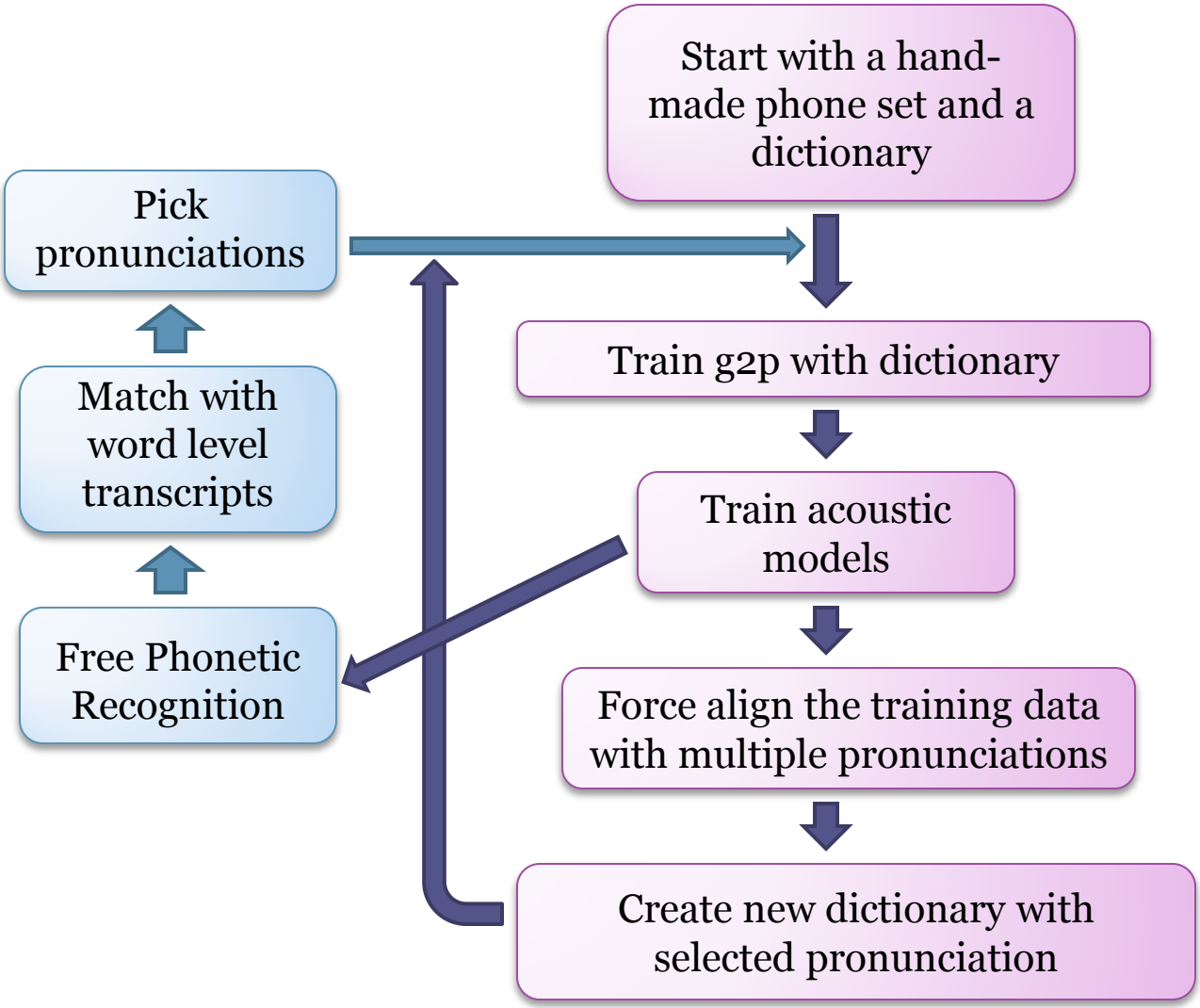
- We use an approximation to the above

$$\hat{Prn} = \arg\max_{Prn \in \{Top\ 5\ Prn\}} P(X \mid Prn)$$

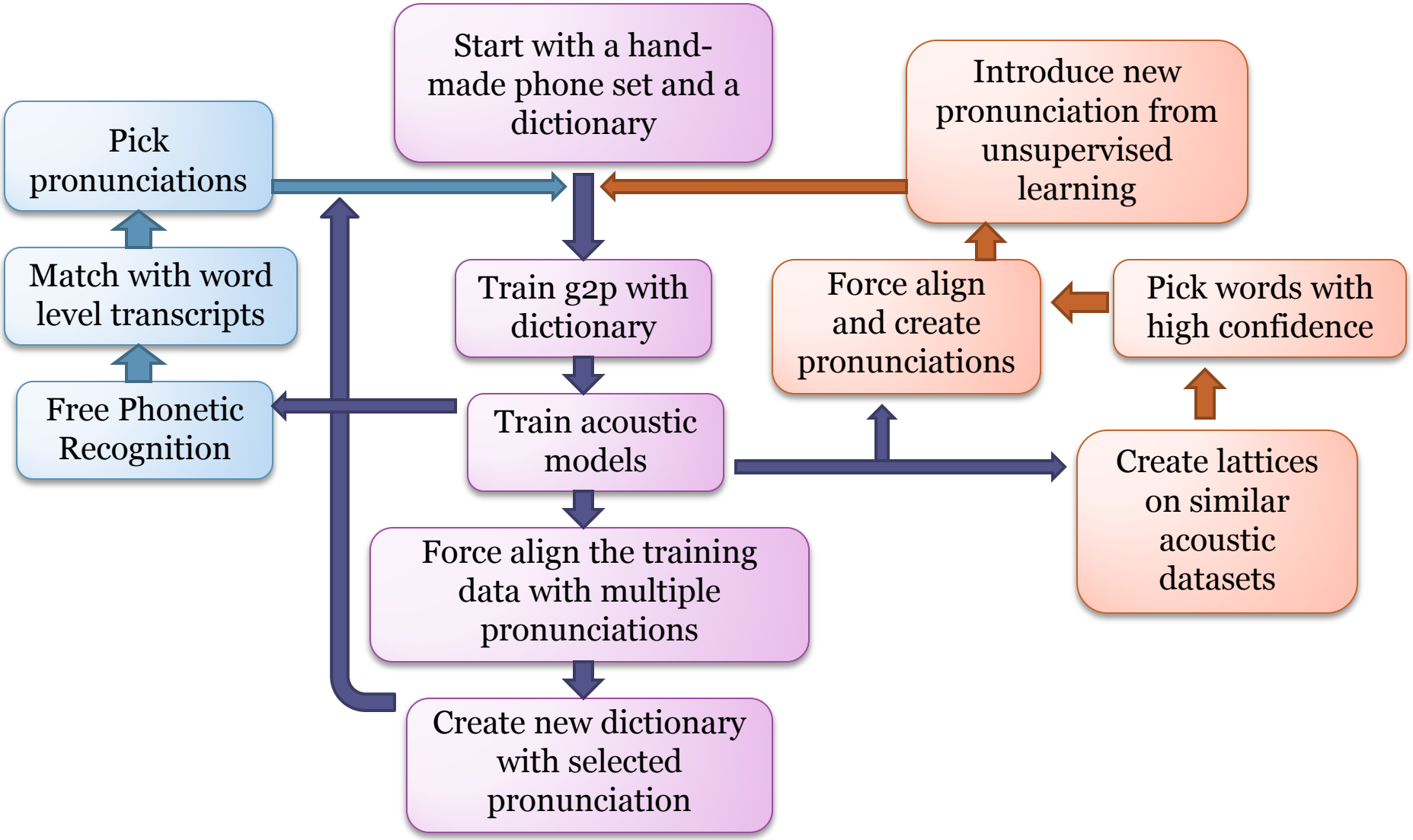
Estimating Pronunciations…



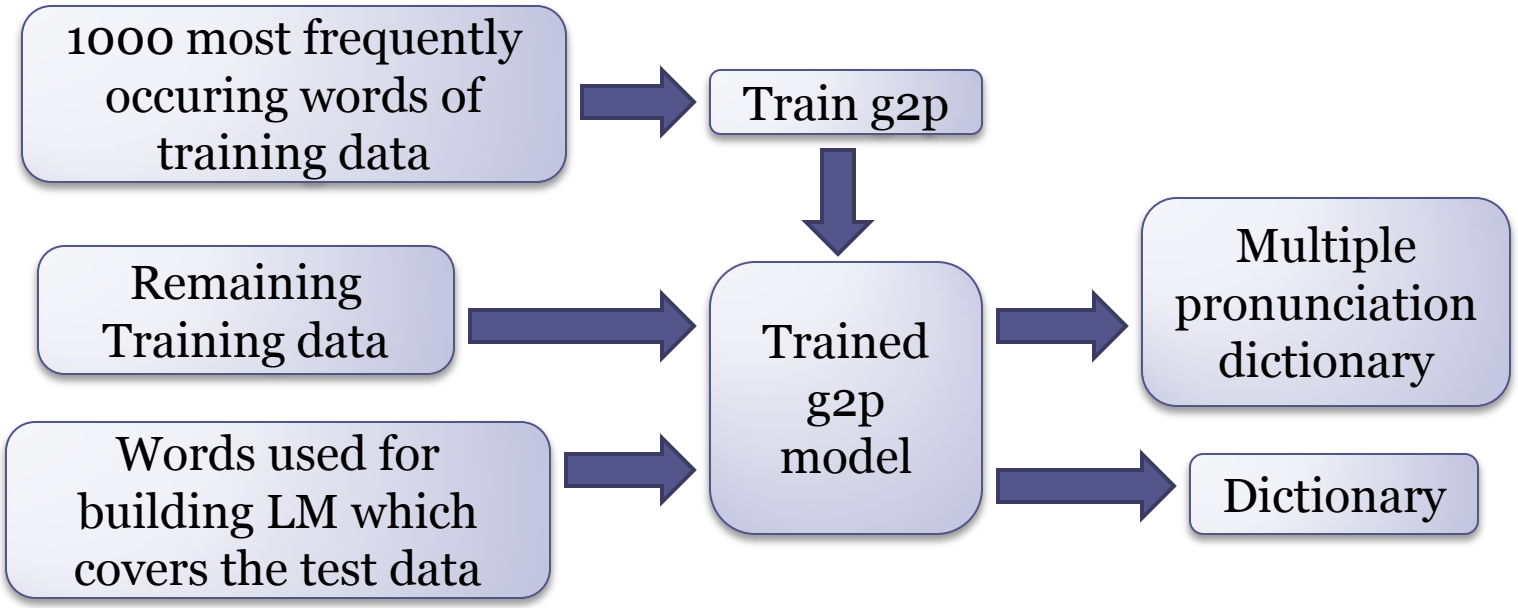
Estimating Pronunciations…



Estimating Pronunciations…

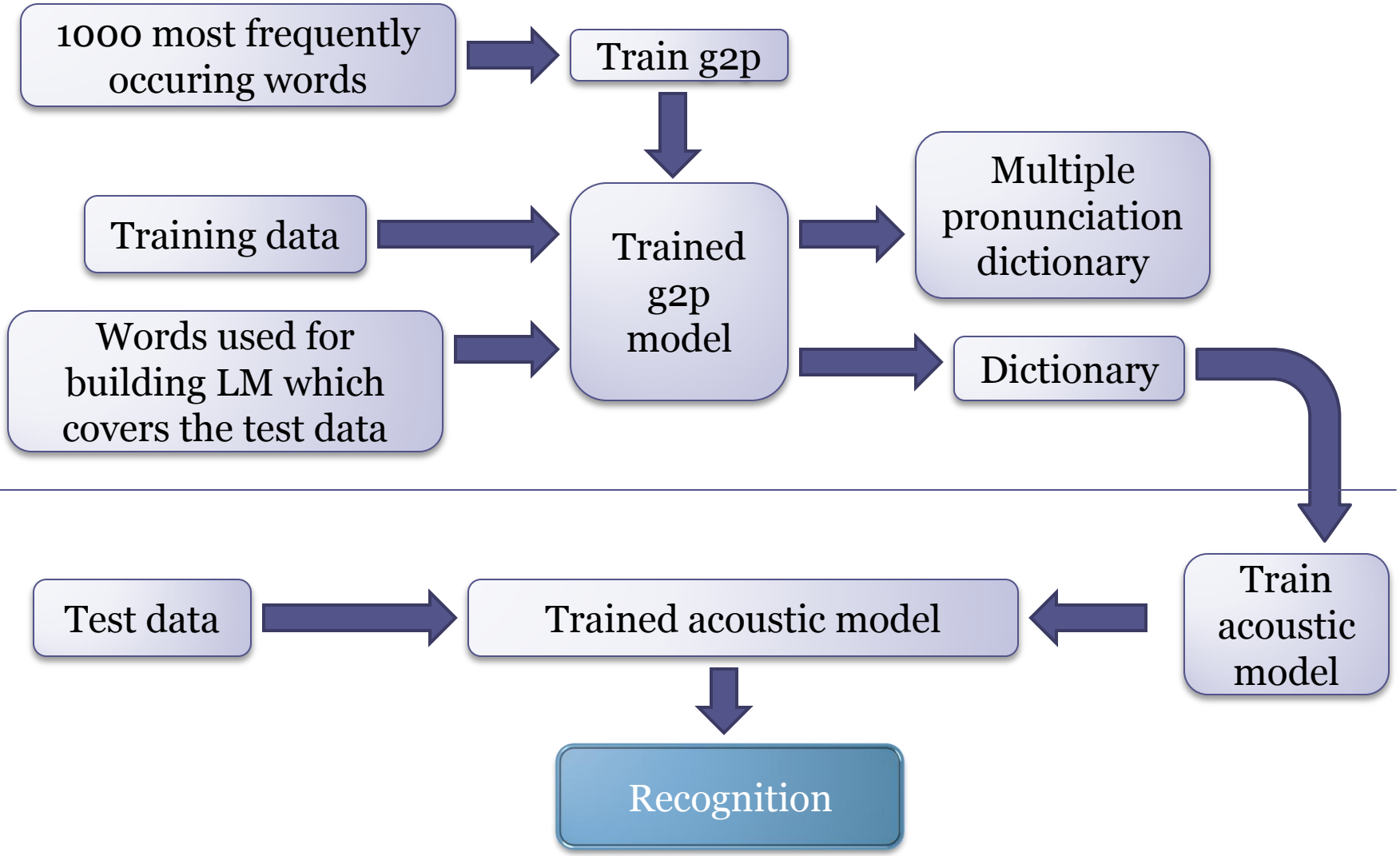


Training Procedure - Bootstrapping

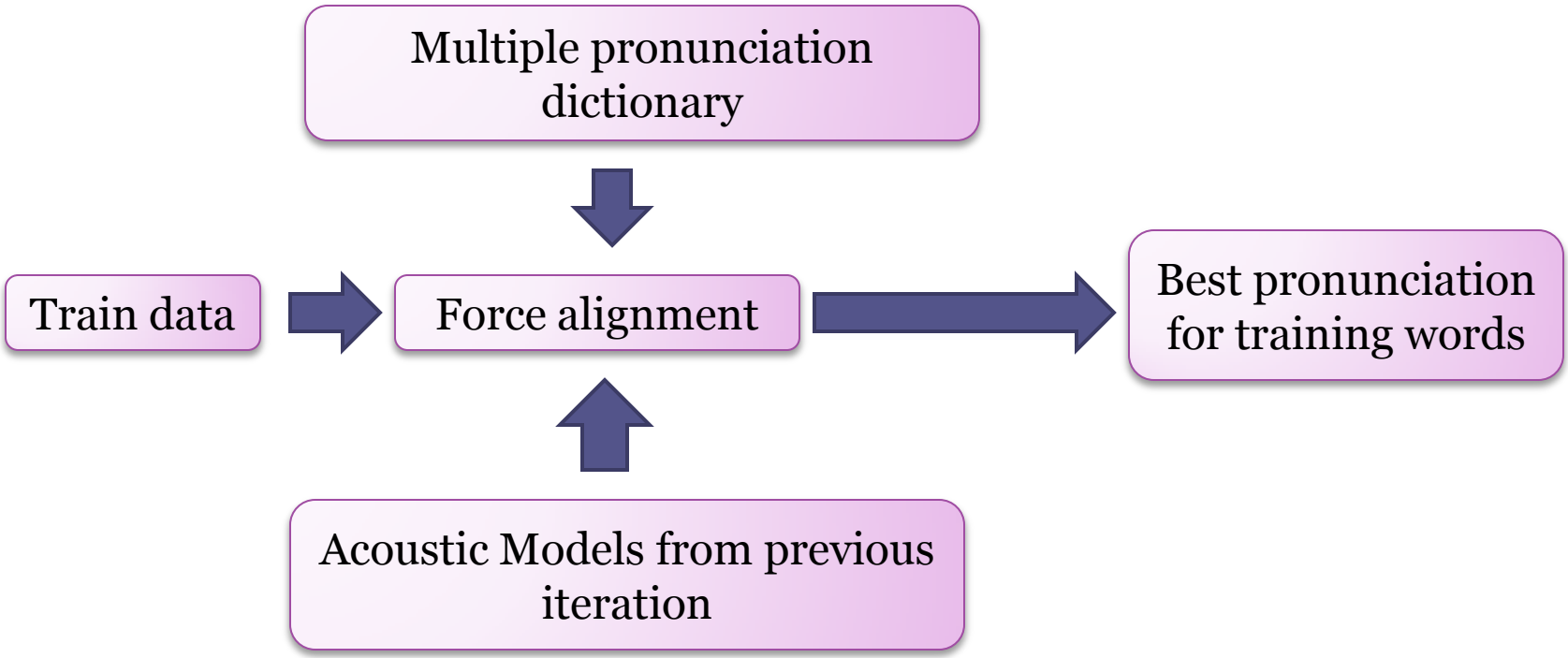


- Callhome training lexicon size – 5 K
- LM vocabulary size – 62 K
- Training acoustic data without partial words – 6 hrs
- Complete training data – 15 hrs

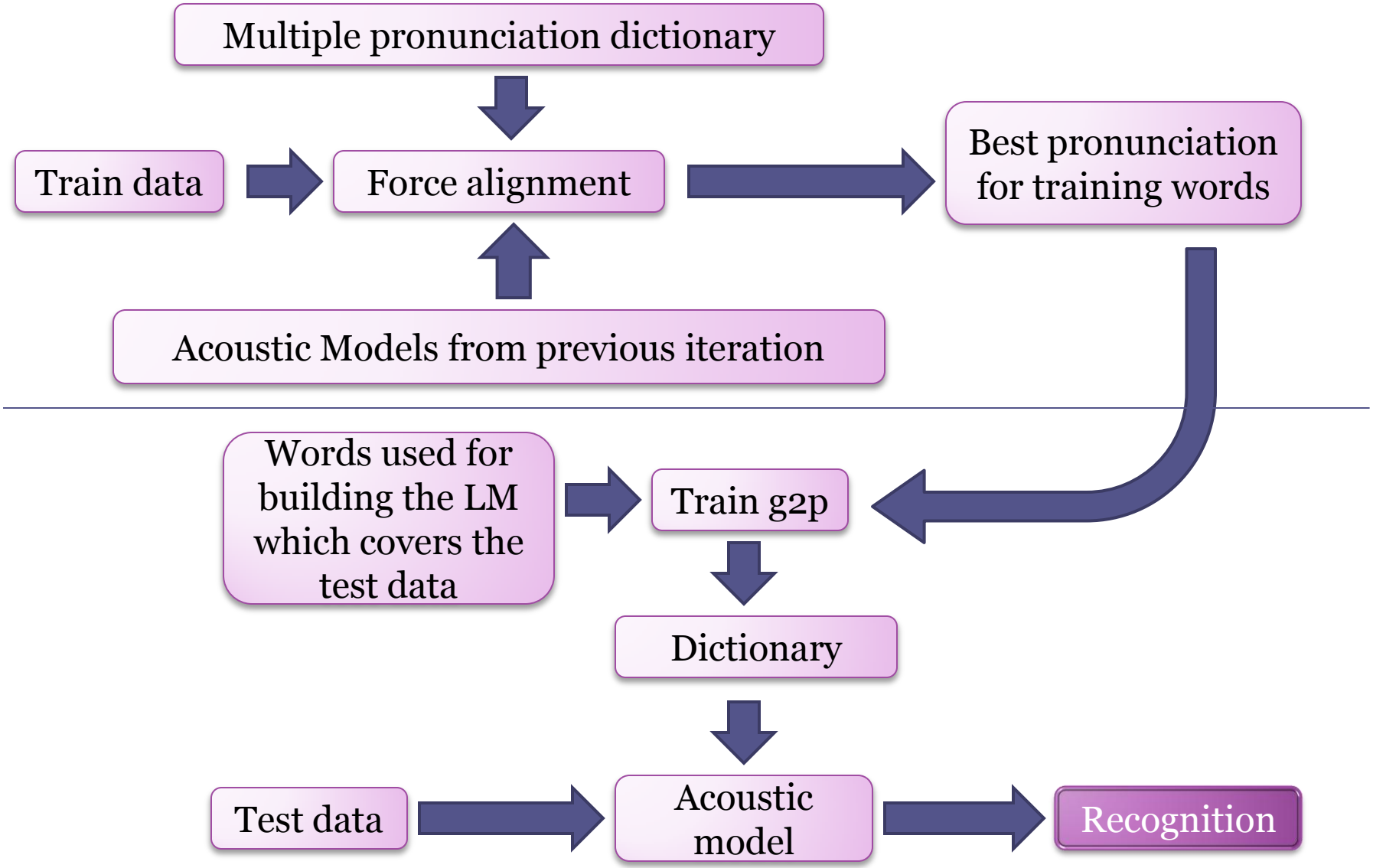
Training Procedure – Bootstrapping



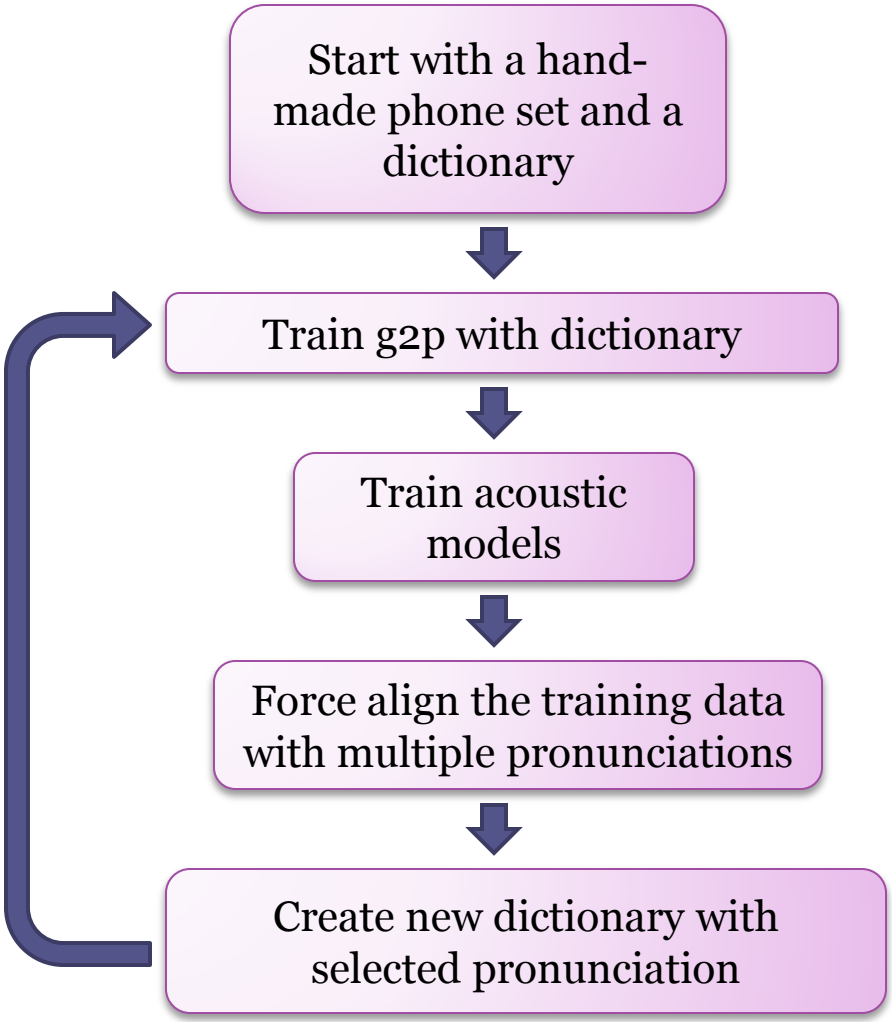
Training Procedure – Building Up



Training Procedure - Building Up



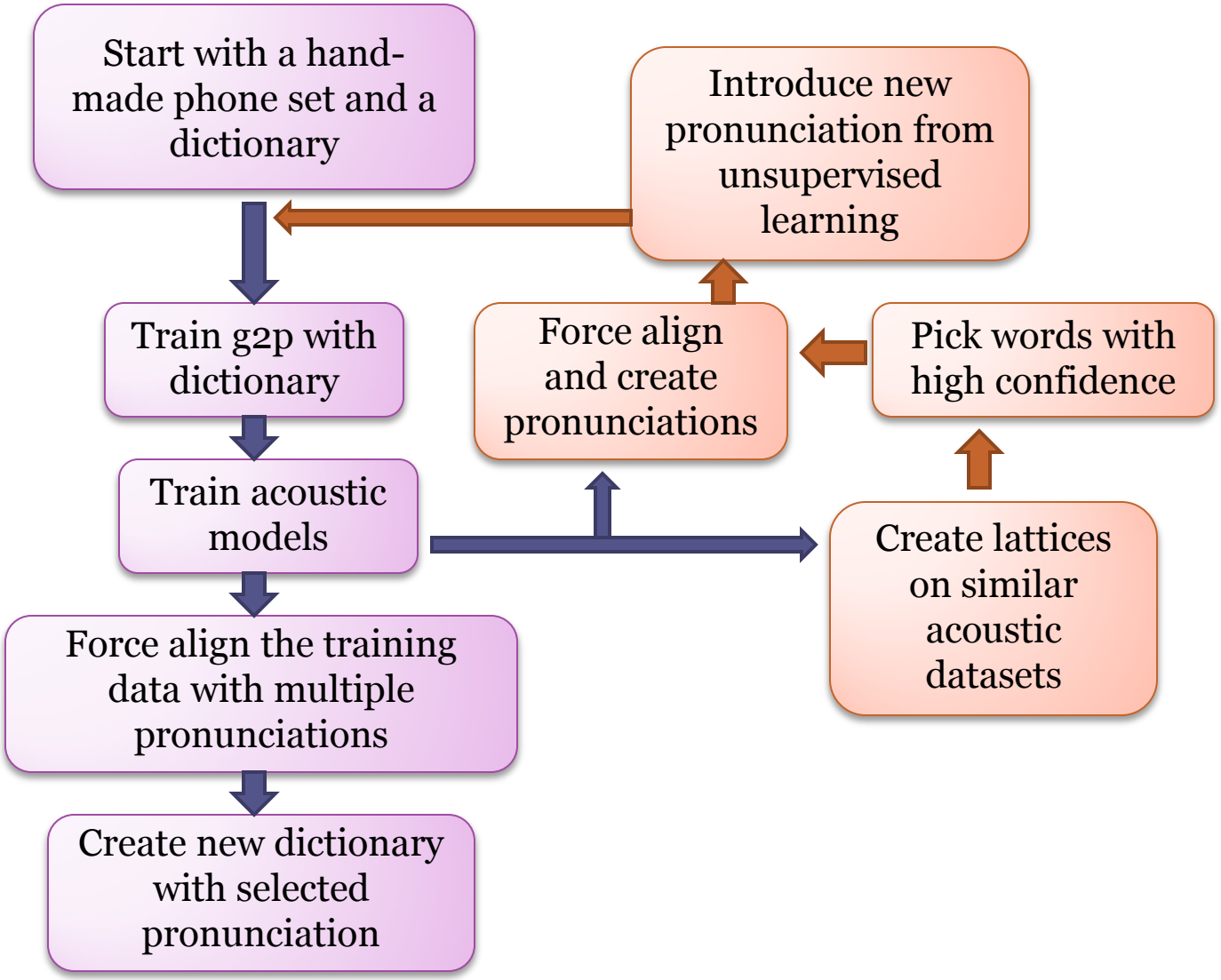
Training Procedure – Building Up



Results

Results	%
Accuracy with full dictionary available	44.35
Accuracy if 5K manual lexicon is available	40.53
Accuracy with 1000 words available	37.58
After retraining acoustic models	39.37
2nd iteration of g2p & acoustic re-train	41.60
3rd iteration of g2p & acoustic re-train	42.11
After increasing the amount of data to 15 hrs	43.56

Unsupervised Learning



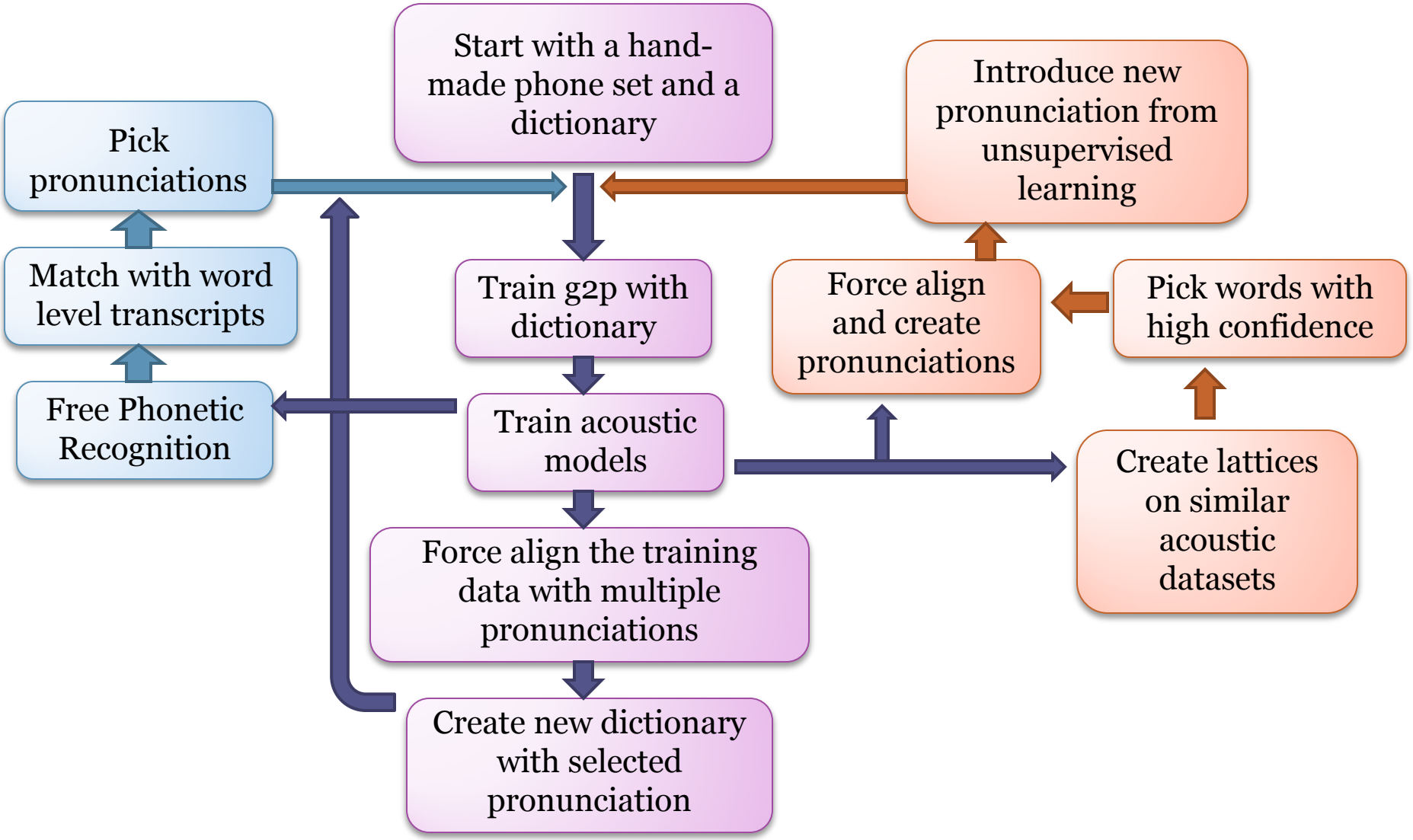
Unsupervised Lexicon Learning Results

	Baseline accuracy	After Unsupervised Learning
6 Hrs of training data	42.11	42.33
15 Hrs of training data	43.56	43.44

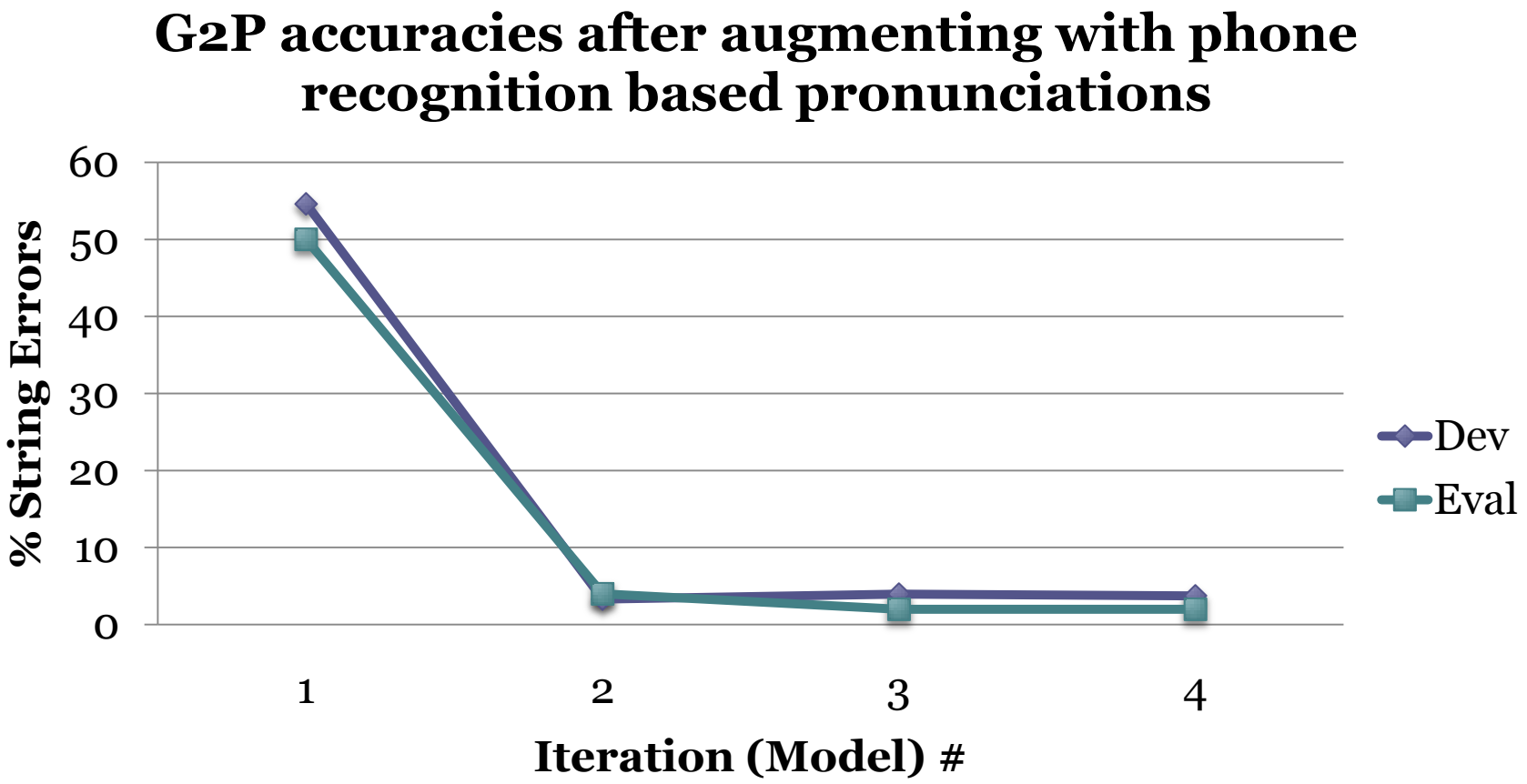
WER dilemma for Spanish Callhome

- Spanish pronunciation is very graphemic
- Accuracy for Spanish are about 31.13% (about 13% lower than callhome english)
- Phone recognition accuracy is better than callhome english
  English: 45.13% Spanish: 53.77%
- LM Perplixity is not too bad: 127
- Can learning alternate pronunciations of *reduced* words help?

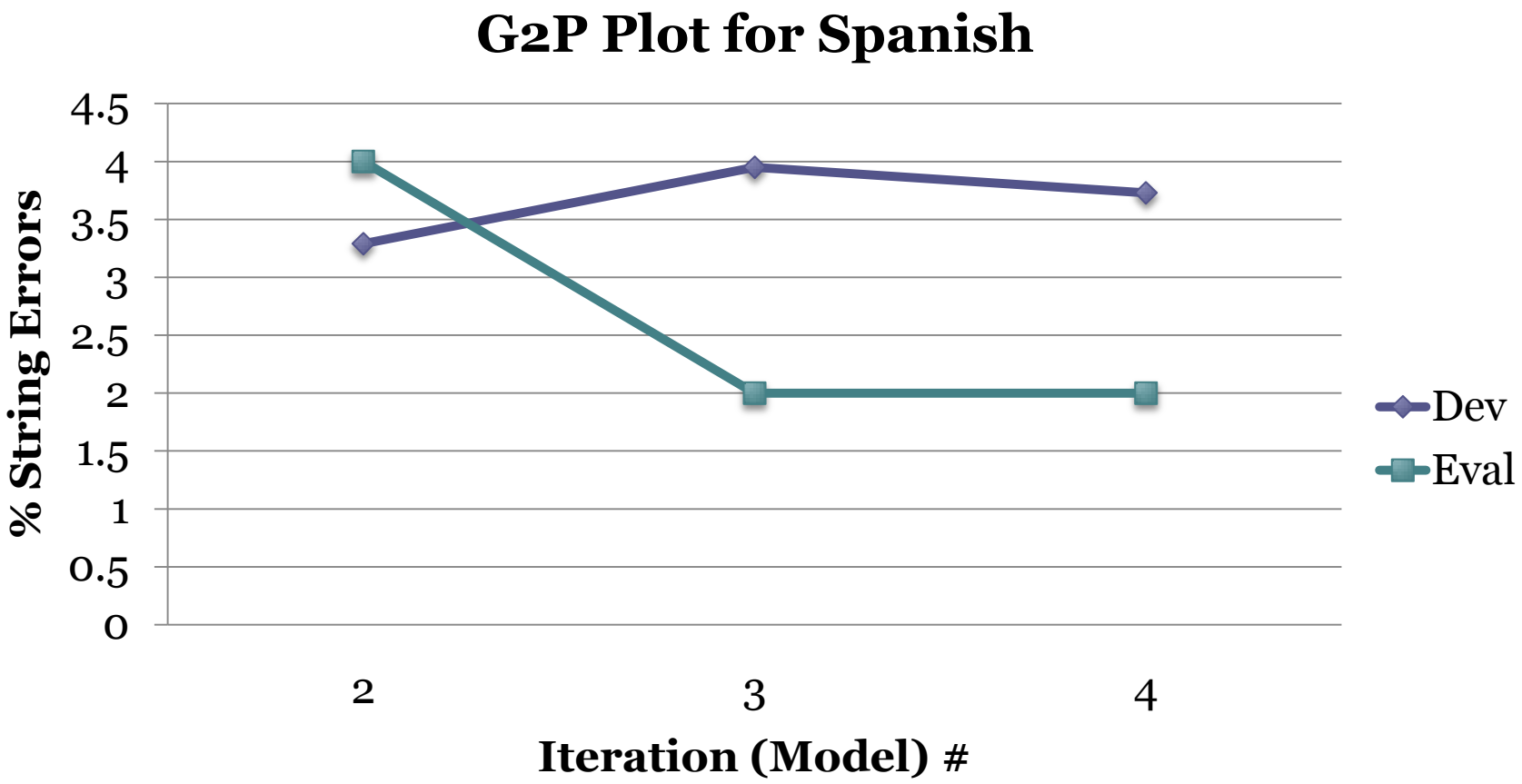
Possible lexicon training paths…



Lexicon Enhancement for Spanish



Lexicon Enhancement for Spanish



English Results and Spanish Results with unconstrained phonetic recognition approach

	Baseline	After adding pronunciations
Spanish	31.13	30.71
English	43.54	42.71

- Log likelihood of training data increases with the new lexicon.

Lexicon Enhancement

- Keep the manual Lexicon but augment with most likely pronunciation in the training data
- Affected about 250 pronunciations

- Accuracy improved from 44.33 to 45.01%

- Multiple Pronunciations had no significant impact: 45.02%

Summary

- G2p based lexicon retraining method helps in achieving accuracies close to hand made lexicons
- It can also help in improving an existing lexicon
- Unsupervised lexicon learning approach and phonetic recognition based lexicon learning approaches hold promise and need to be explored with a wider variety of smoothing and pronunciation extraction scenarios

Training Procedure

- Train g2p to generate pronunciations using your best baseline lexicon

- Generate multiple pronunciations using the g2p

- Use the training data to select the best pronunciation out of these multiple choices

- Retrain the acoustic models and iterate over the above process