Overcoming Resource Limitations in the Processing of Unlimited Speech:

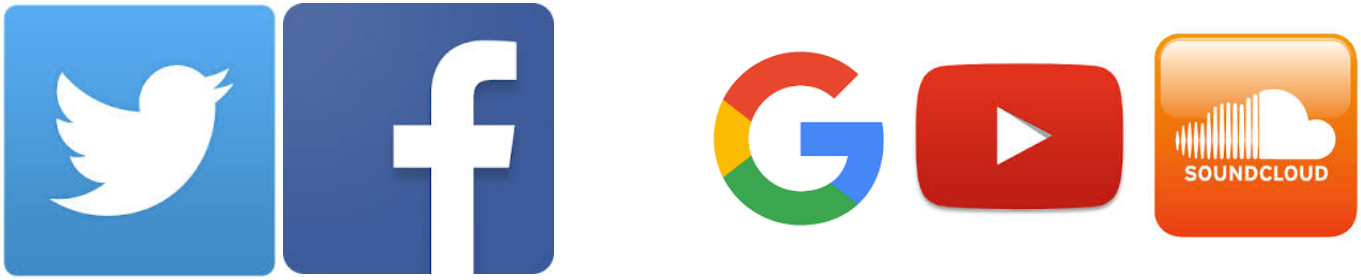
Applications to Speaker and Language Recognition

Stephen H. Shum

4 May 2016

Motivation

- Unlimited access to data

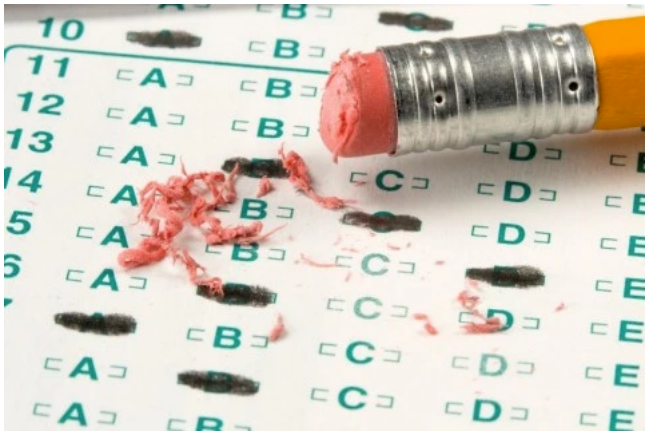


- Limited access to proper tagging and annotation

Motivation

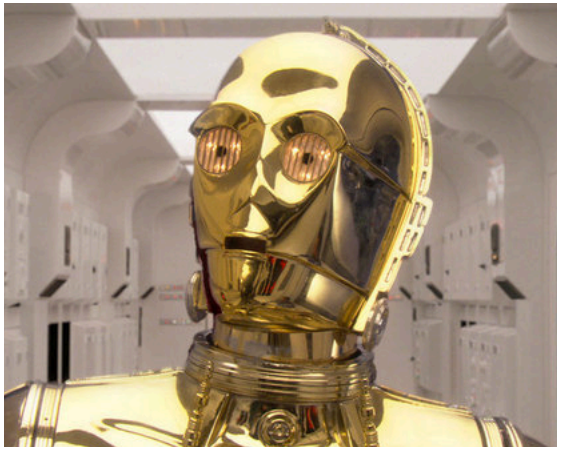
- Need for adaptation

CollegeBoard
achieve more
SAT



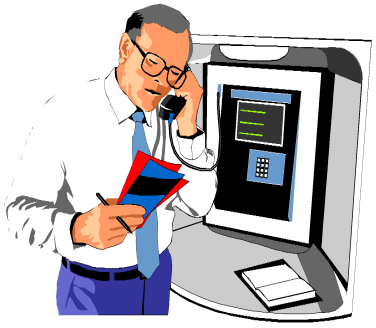
- Hard to scale!

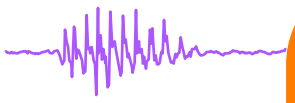
Perspective



- C-3PO
  - > 6 million forms of communication
- World
  - > 7,000 living languages
  - ~400 languages with > 1 million speakers
- Speech technology
  - < 100 languages
- Thesis
  - < 30 languages

Extracting Information from Speech





Speech Signal

Speech Recognition → Words
"How are you?"

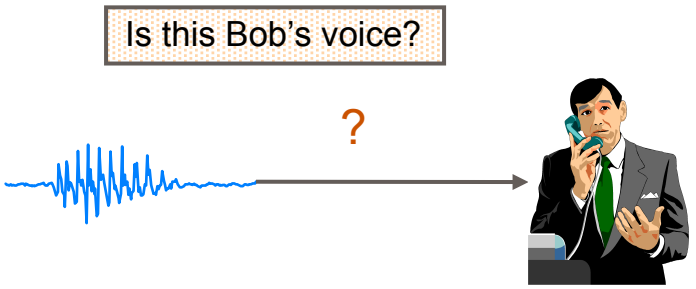
Language Recognition → Language Name
English

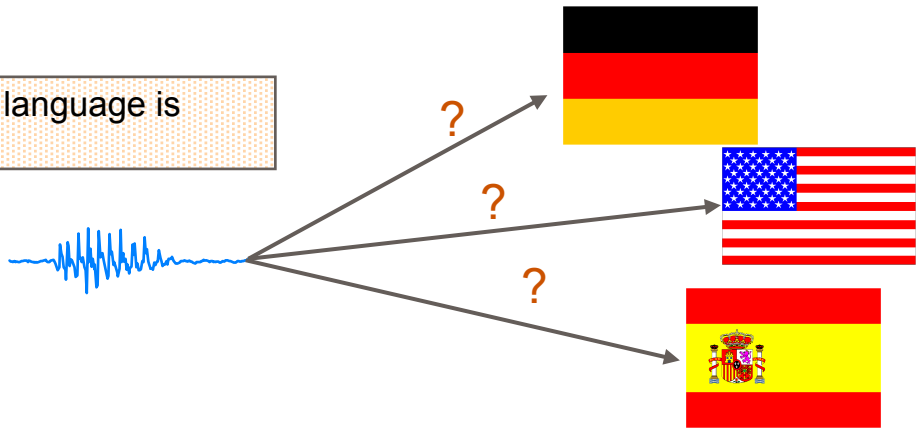
Speaker Recognition → Speaker Name
James Wilson

Emotion Recognition → Emotion category
Happy, angry …

Tasks

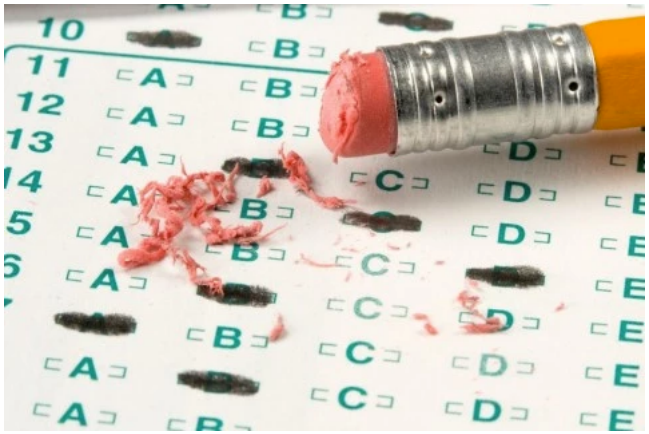
- Speaker verification
  - Determine whether or not a test utterance was spoken by a particular speaker
- Language identification
  - Determine, from a known set of target languages, the spoken language of a test utterance





Themes

- Domain adaptation



Supervised domain adaptation!

Themes

- Unsupervised domain adaptation
  - Access to many test preparation resources,
  - but **no** access to their answer keys!

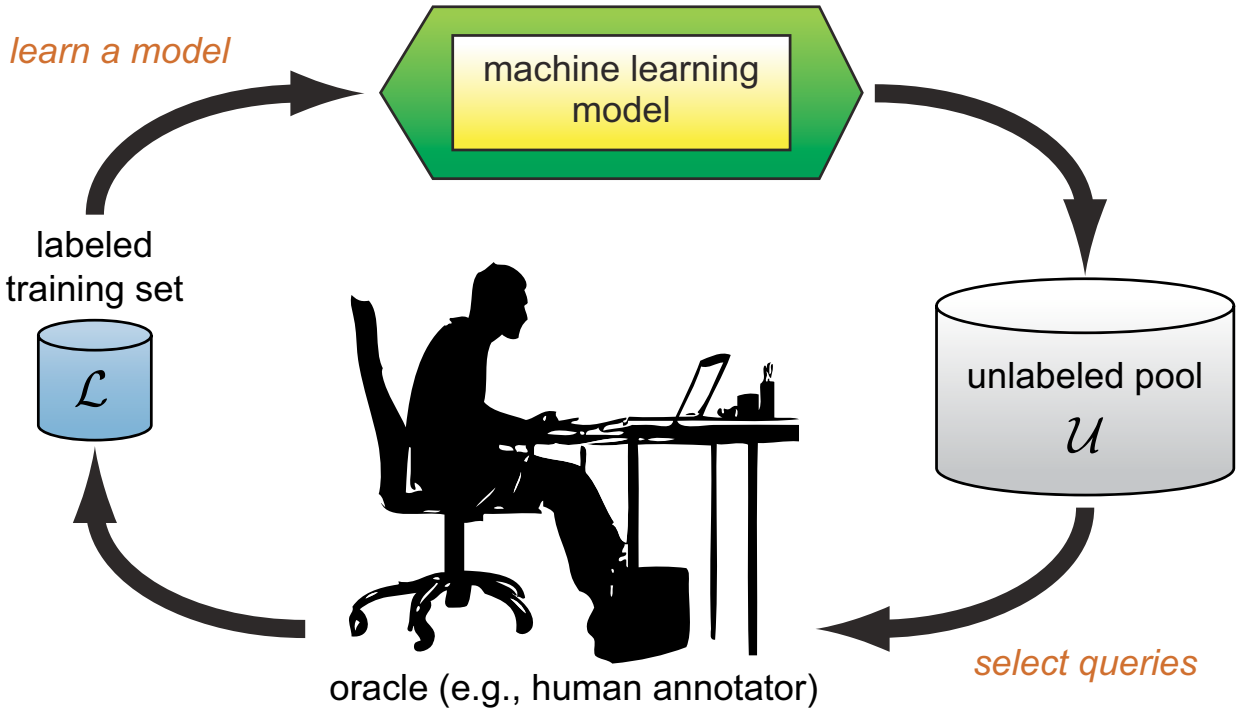
  - Tests and test conditions change continuously;
  - we'd like to be able to adapt to these changes without needing a new study guide (and corresponding answer key!) every time.

Overview

	Domain Adaptation	
Speaker Verification	Adapt system to changes in recording technology by applying existing models to new, unlabeled data sets	
Language Identification	Augment existing volumes of transcribed speech with large-scale, unsupervised discovery of acoustic units on untranscribed, multilingual data	

Themes

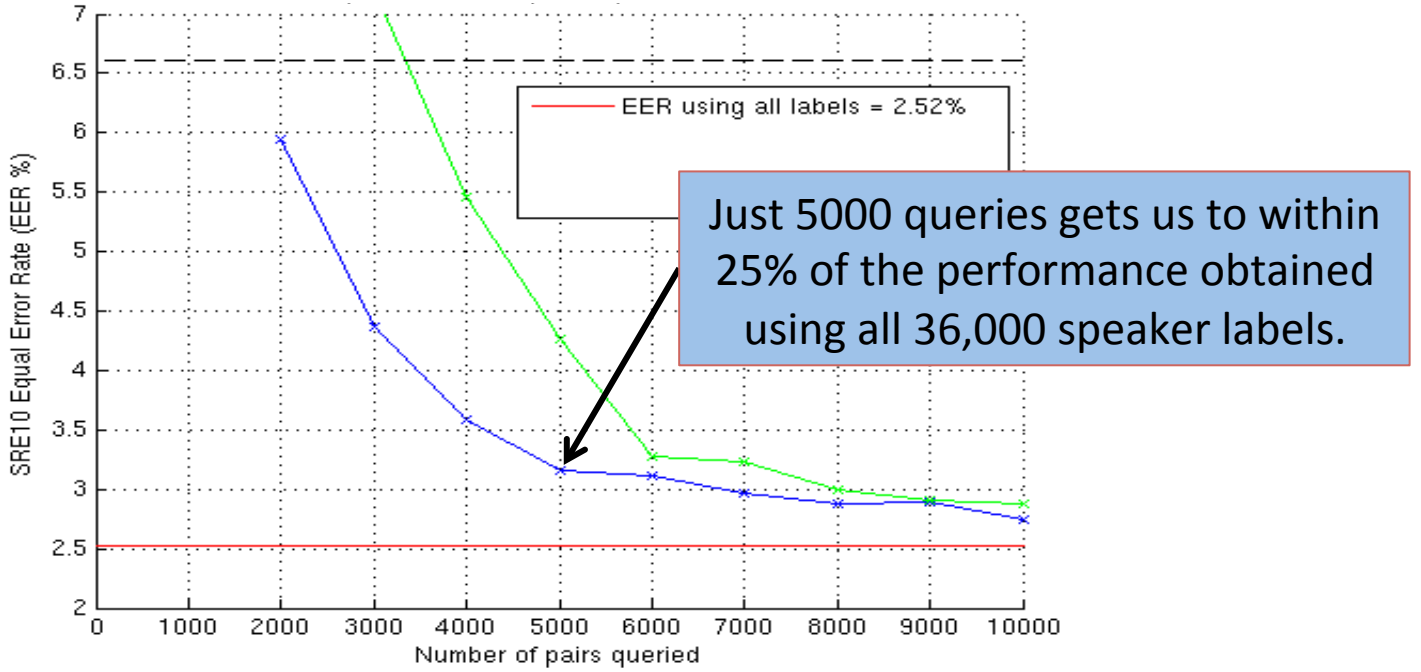
- Learning from weak supervision
  - Active learning



Burr Settles, "Active Learning Literature Survey," Computer Sciences Technical Report 1648, University of Wisconsin-Madison, 2010.

Themes

- Learning from weak supervision
  - Active learning
    - Choosing what gets labeled yields a dramatic reduction in the number of labels needed to achieve desired performance



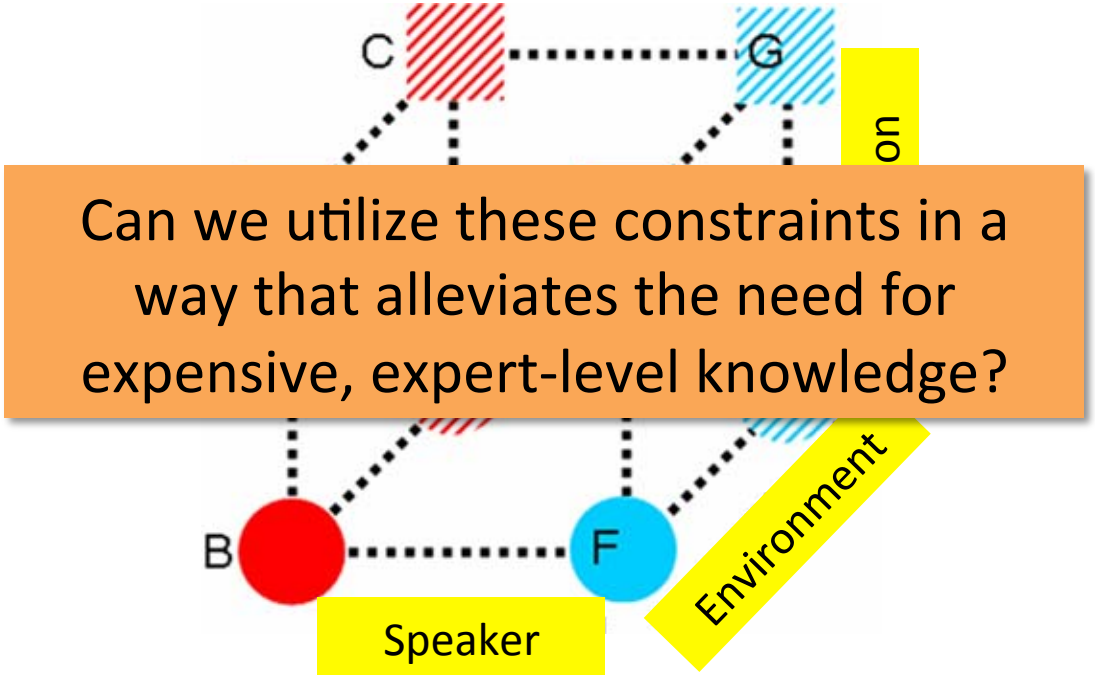
S. Shum, N. Dehak, J. Glass, "Limited Labels for Unlimited Data: Towards Active Learning for Speaker Recognition," *Proceedings of Interspeech*, 2014.

Overview

	Domain Adaptation	Weak Supervision
Speaker Verification	Adapt system to changes in recording technology by applying existing models to new, unlabeled data sets	Actively explore a database of unknown speakers and build speaker models using pairwise equivalence constraints
Language Identification	Augment existing volumes of transcribed speech with large-scale, unsupervised discovery of acoustic units on untranscribed, multilingual data	

Themes

- Learning from weak supervision
  - Active learning
  - Top-down equivalence constraints



R. Hammer, T. Hertz, Shaul Hochstein, and Daphna Weinshall, "Category learning from equivalence constraints," *Cognitive Processing*, 2008.

Overview

	Domain Adaptation	Weak Supervision
Speaker Verification	Adapt system to changes in recording technology by applying existing models to new, unlabeled data sets	Actively explore a database of unknown speakers and build speaker models using pairwise equivalence constraints
Language Identification	Augment existing volumes of transcribed speech with large-scale, unsupervised discovery of acoustic units on untranscribed, multilingual data	Use equivalence constraints between acoustic sequences to improve speaker-independence of discovered acoustic units

Overview

	Domain Adaptation	Weak Supervision
Speaker Verification	Adapt system to changes in recording technology by applying existing models to new, unlabeled data sets	Actively explore a database of unknown speakers and build speaker models using pairwise equivalence constraints
Language Identification	Augment existing volumes of transcribed speech with large-scale, unsupervised discovery of acoustic units on untranscribed, multilingual data	Use equivalence constraints between acoustic sequences to improve speaker-independence of discovered acoustic units

Domain Adaptation for Speaker Recognition

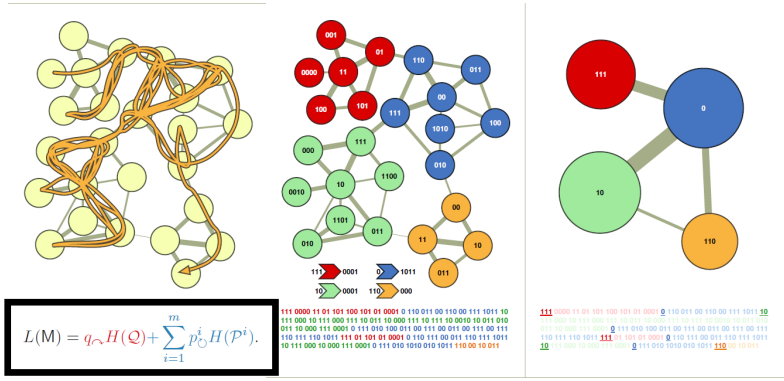
Labeled Data	Unlabeled Data	EVAL Result
SWB	none	5.54%
SRE	none	2.30%

Results shown are the Equal Error Rate (EER) and can be interpreted as a measure of error.

- Caused by changes in recording technology?
  - SWB collected from 1992 – 2000 (mostly landline)
  - SRE collected from 2004 – 2008 (mostly cellular)
  - EVAL collected in 2010

Domain Adaptation for Speaker Recognition

- Challenge Task
  - SWB with speaker labels
  - SRE without speaker labels
  - Evaluate on EVAL



- Proposed "bootstrap" framework
  - Use labeled data to model unlabeled data
  - Cluster unlabeled data using a combination of
    - random walk-based graph clustering (Infomap)
    - agglomerative hierarchical clustering
  - Interpolate between resulting hyper-parameters

Proposed (Bootstrap) Framework

- Use speaker labels and SWB to obtain $\{\Sigma, \Phi\}_{\text{SWB}}$
- Use $\{\Sigma, \Phi\}_{\text{SWB}}$ to represent SRE in the form of a pairwise affinity matrix, A
- Cluster A to obtain (hypothesized) speaker labels for SRE
- Use speaker labels and SRE obtain $\{\Sigma, \Phi\}_{\text{SRE}}$
- Linearly interpolate (via $\{\alpha_{\text{WC}}, \alpha_{\text{AC}}\}$) between $\{\Sigma, \Phi\}_{\text{SWB}}$ and $\{\Sigma, \Phi\}_{\text{SRE}}$ to obtain

$$\Sigma_{\text{F}} = \alpha_{\text{WC}} \cdot \Sigma_{\text{SRE}} + (1 - \alpha_{\text{WC}}) \cdot \Sigma_{\text{SWB}}$$

$$\Phi_{\text{F}} = \alpha_{\text{AC}} \cdot \Phi_{\text{SRE}} + (1 - \alpha_{\text{AC}}) \cdot \Phi_{\text{SWB}}$$

Domain Adaptation for Speaker Recognition

Labeled Data	Unlabeled Data	EVAL Result
SWB	none	5.54%
SWB	SRE	2.53%
SRE	none	2.30%
SWB + SRE	none	2.23%

Results shown are the Equal Error Rate (EER) and can be interpreted as a measure of error.

- Our proposed adaptation system achieves EVAL performance that is within 15% of a system that has access to all speaker labels.

S. Shum, D. Reynolds, D. Garcia-Romero, and A. McCree, "Unsupervised Clustering Approaches for Domain Adaptation in Speaker Recognition Systems," *Proceedings of Odyssey*, 2014. [Best Student Paper]

Overview

	Domain Adaptation	Weak Supervision
Speaker Verification	Adapt system to changes in recording technology by applying existing models to new, unlabeled data sets	Actively explore a database of unknown speakers and build speaker models using pairwise equivalence constraints
Language Identification	Augment existing volumes of transcribed speech with large-scale, unsupervised discovery of acoustic units on untranscribed, multilingual data	Use equivalence constraints between acoustic sequences to improve speaker-independence of discovered acoustic units

Overview

	Domain Adaptation	Weak Supervision
Speaker Verification	Adapt system to changes in recording technology by applying existing models to new, unlabeled data sets	Actively explore a database of unknown speakers and build speaker models using pairwise equivalence constraints
Language Identification	Augment existing volumes of transcribed speech with large-scale, unsupervised discovery of acoustic units on untranscribed, multilingual data	Use equivalence constraints between acoustic sequences to improve speaker-independence of discovered acoustic units

Acoustic Unit Discovery for Language Identification

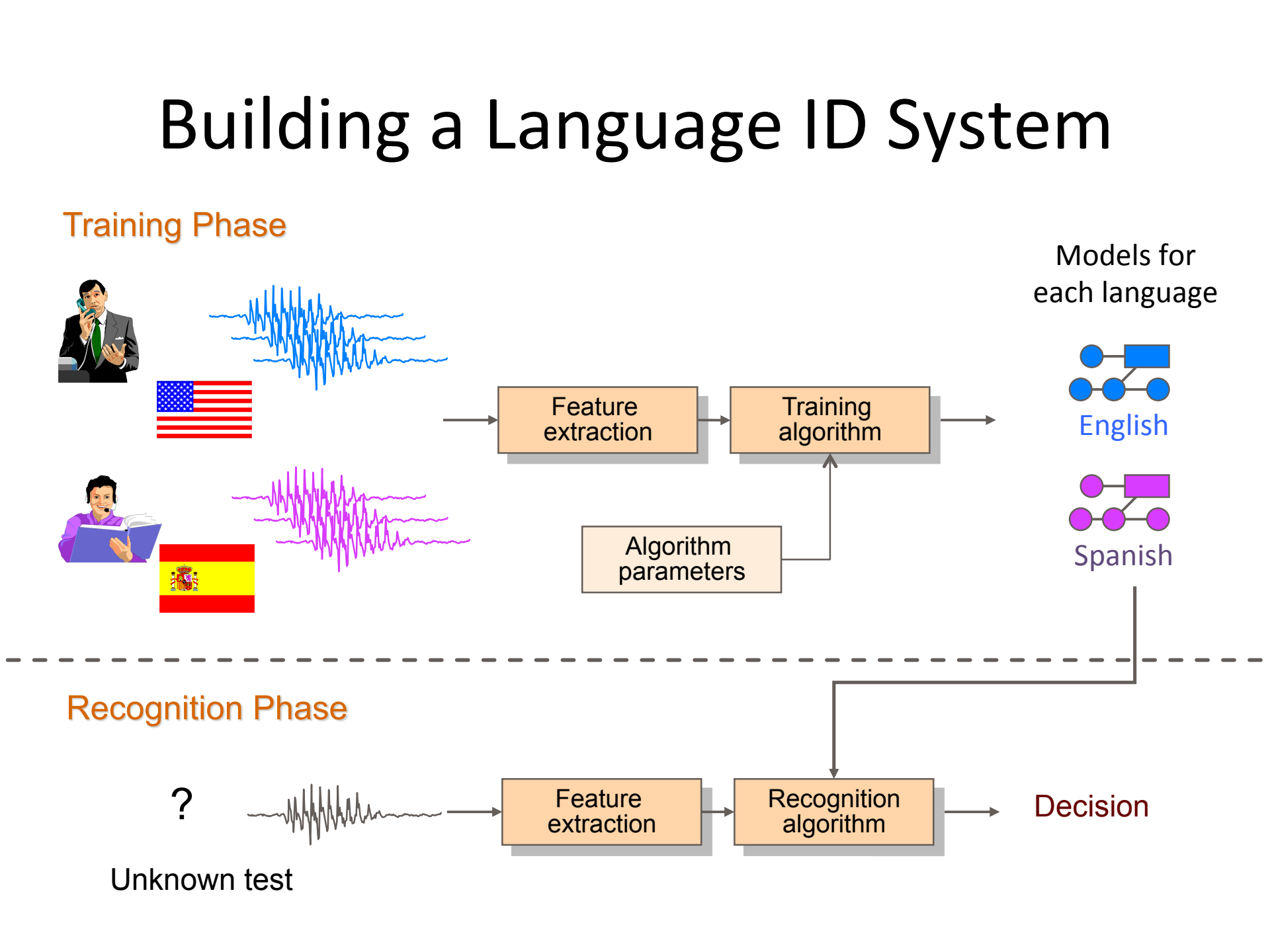
- Language recognition using i-vectors
  - Spectral feature baseline
  - DNN bottleneck feature benchmark
- Parallelizing a Bayesian nonparametric model for large-scale acoustic unit discovery
- Experiments
  - The usefulness of context-dependent modeling
  - The magic of fusion
  - The impact of improved acoustic features
  - The generalizability of language-specific perspectives

NIST Language Recognition Evaluation 2011

- 24 languages
  - Arabic (Iraqi, Levantine, Maghrebi, MSA), Bengali, Czech, Dari, English (American, Indian), Farsi, Hindi, Laotian, Mandarin, Pashto, Polish, Punjabi, Russian, Slovak, Spanish, Tamil, Thai, Turkish, Ukrainian, Urdu

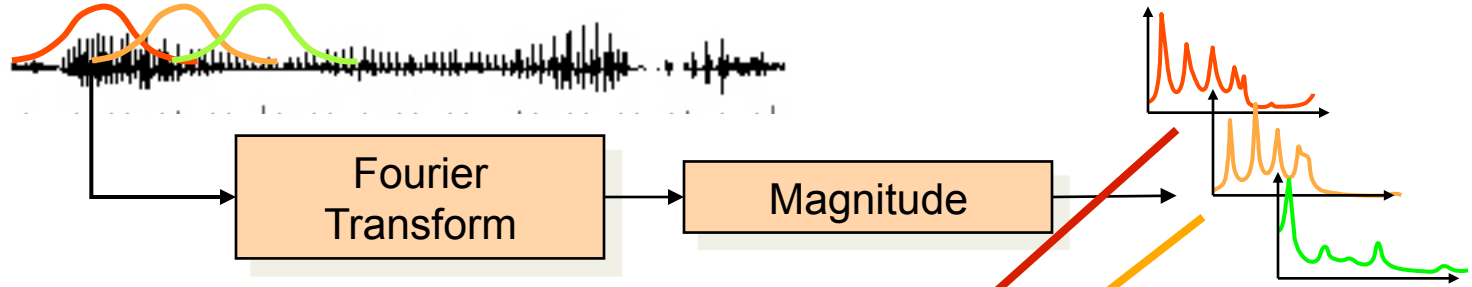
- Identify language from 30s / 10s / 3s segments

Building a Language ID System

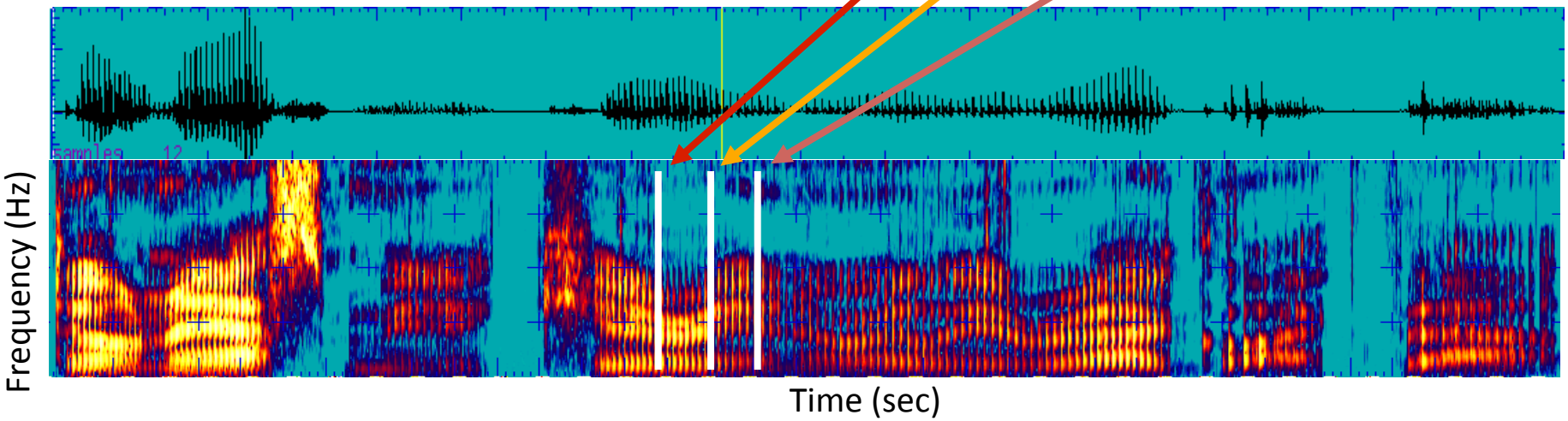


Overview of Spectral Features

- We capture speech information via a time sequence of spectral features (100 / second)

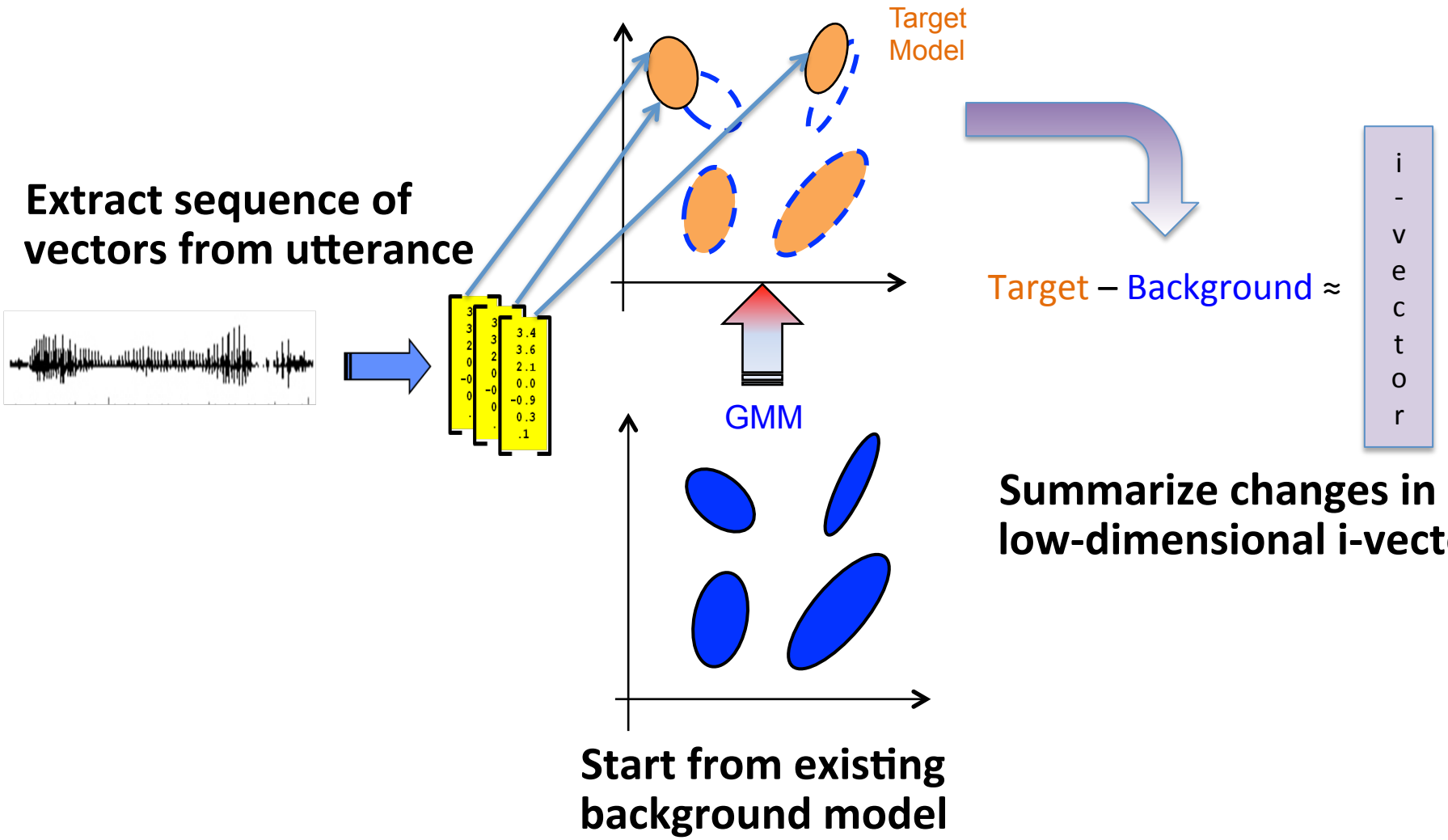


- And produce a time-frequency evolution of the spectrum



The i-vector approach

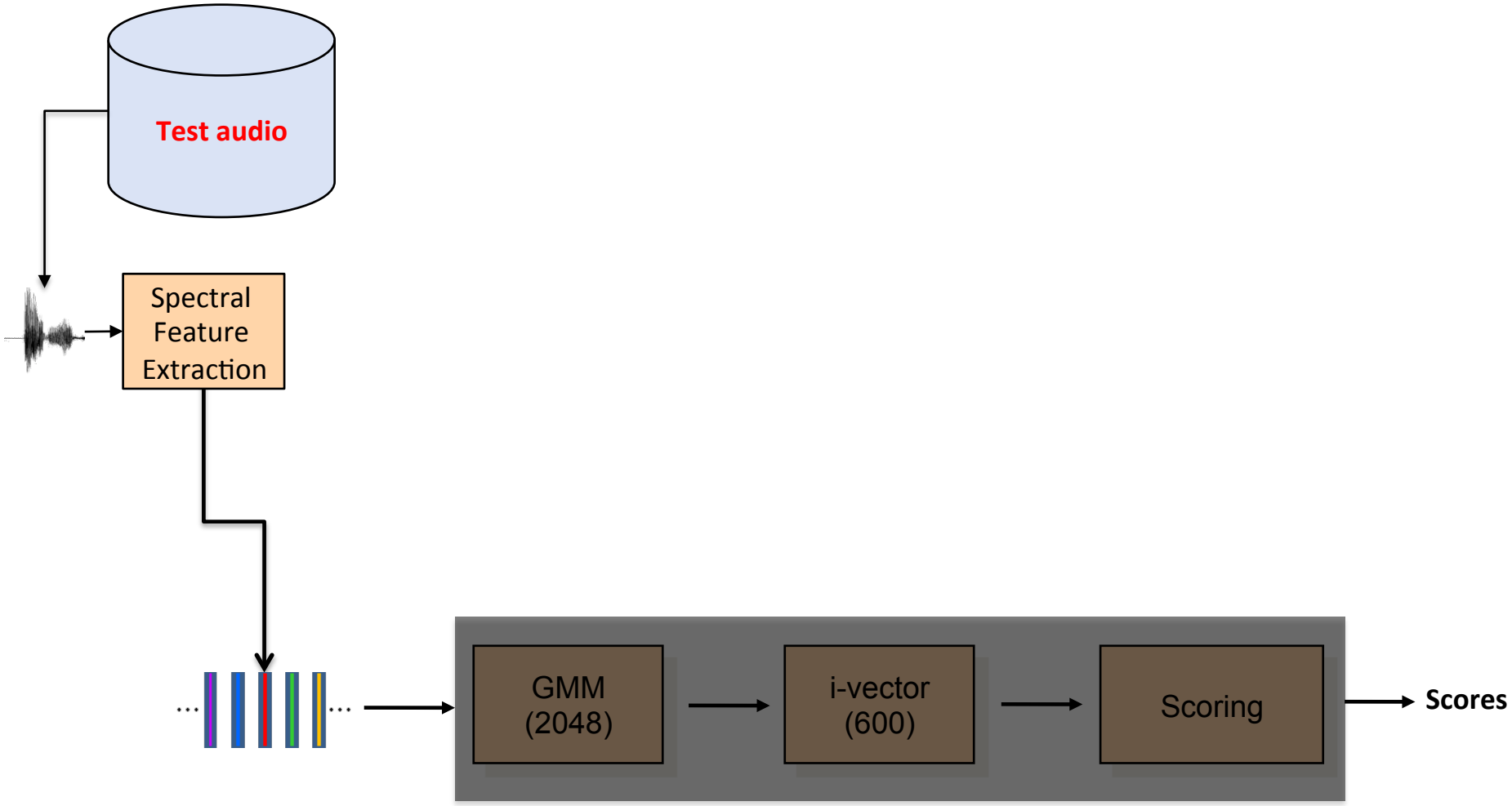
Adapt to obtain target model



Notes about i-vectors

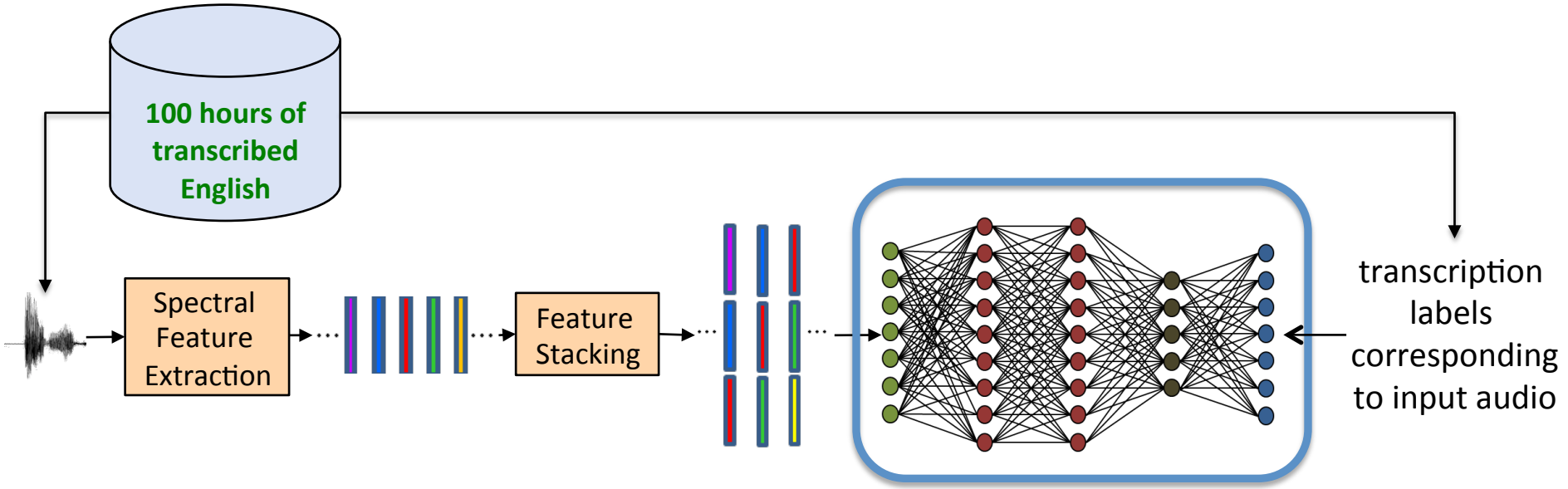
- Utterance length-independent, low-dimensional summary representation of audio
- Not particularly informative by themselves
- Convenient for incorporating information from labeled data

Acoustic i-vector system for language recognition (baseline)



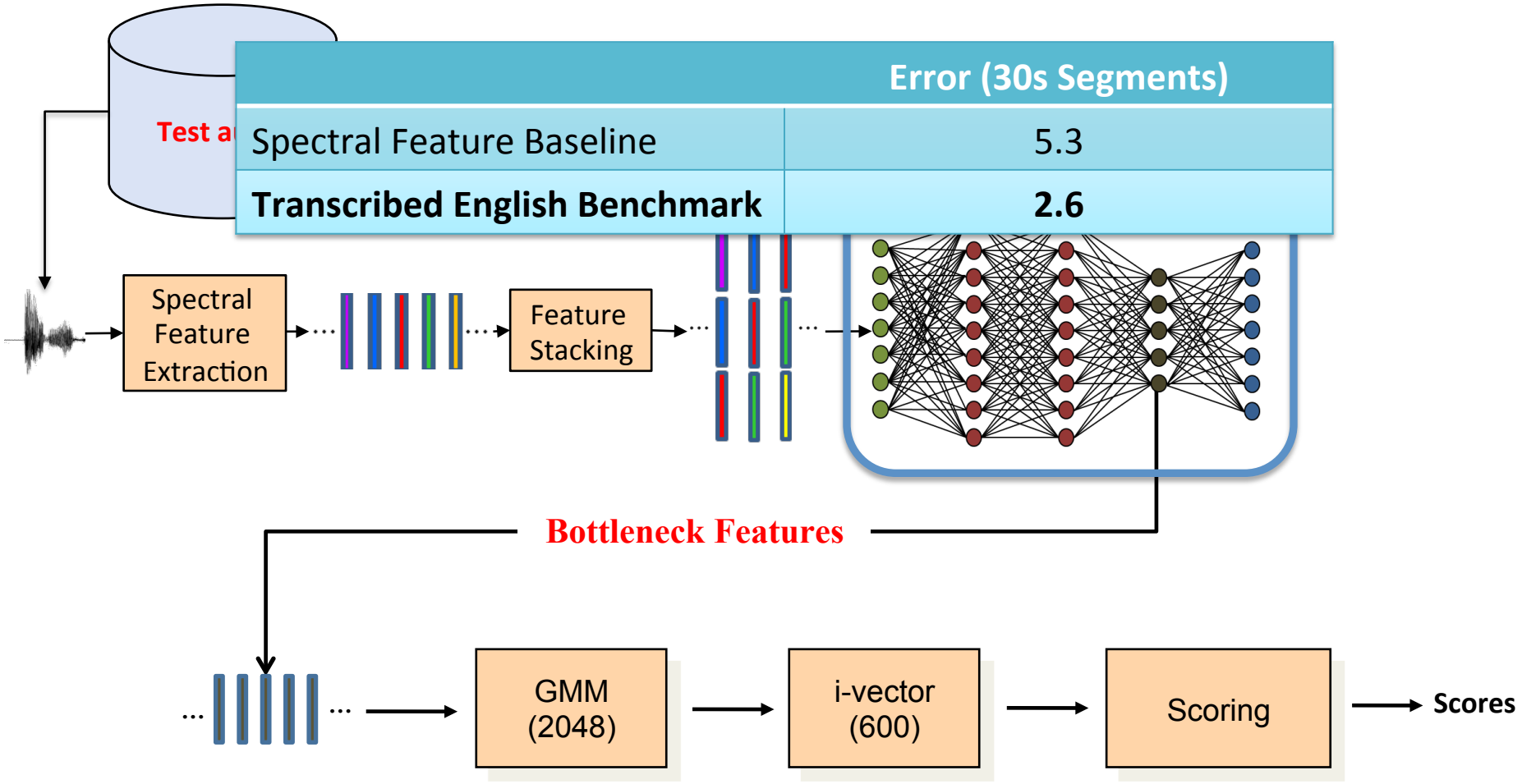
N. Dehak, P. Torres-Carrasquillo, D. Reynolds, and R. Dehak, "Language Recognition via i-vectors and dimensionality reduction", *Proc. Interspeech*, 2011.

Incorporating transcribed English



F. Richardson, D. Reynolds, and N. Dehak, "Deep Neural Network Approaches to Speaker and Language Recognition", *IEEE Signal Processing Letters*, 2015.

Transcribed English-based bottleneck i-vector system (benchmark)



F. Richardson, D. Reynolds, and N. Dehak, "Deep Neural Network Approaches to Speaker and Language Recognition", *IEEE Signal Processing Letters*, 2015.

Why this works

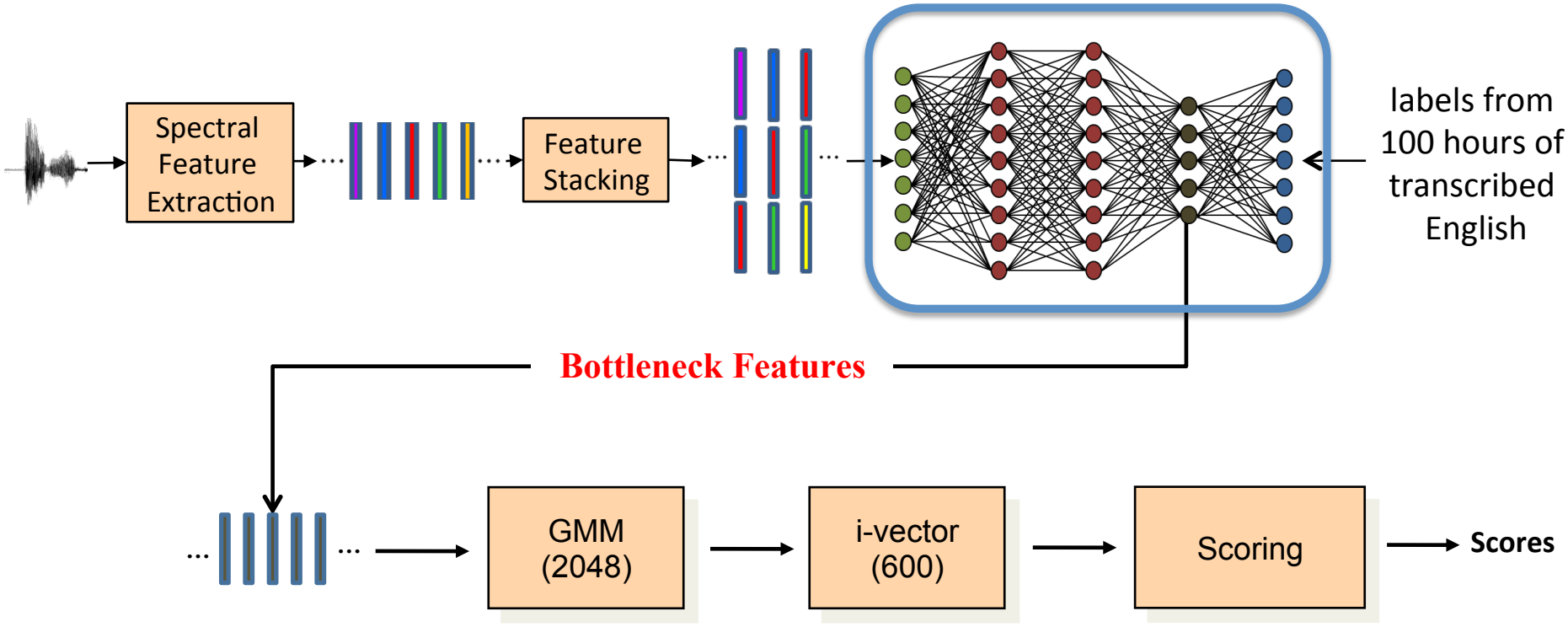
- Increasing phonetic awareness
  - Language comes naturally to humans
  - Analogy
    - A computer identifying spoken language **without** phonetic awareness (i.e., from spectral features)
    - A human identifying birds by their respective song

Comments

- Incorporating transcribed English effectively cuts the error rate in half.
  - But there are 23 other languages!
- Incorporating transcribed data from other languages helps even more.
  - Can we make good use of untranscribed data?

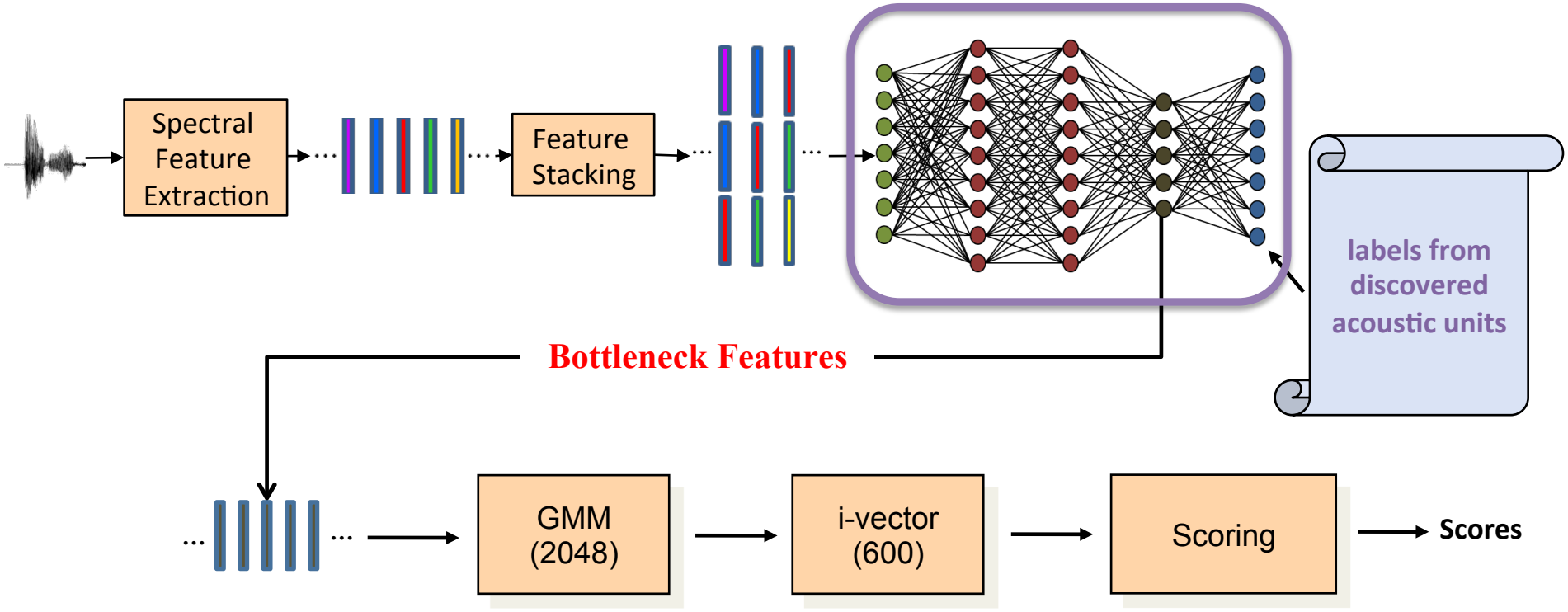
R. Fer, P. Matejka, F. Grezl, O. Plchot, and J. Cernocky, "Multilingual Bottleneck Features for Language Recognition," *Proceedings of Interspeech*, 2015.

Transcribed English-based bottleneck i-vector system (benchmark)



F. Richardson, D. Reynolds, and N. Dehak, "Deep Neural Network Approaches to Speaker and Language Recognition", *IEEE Signal Processing Letters*, 2015.

Acoustic unit discovery-based bottleneck i-vector system (proposed)

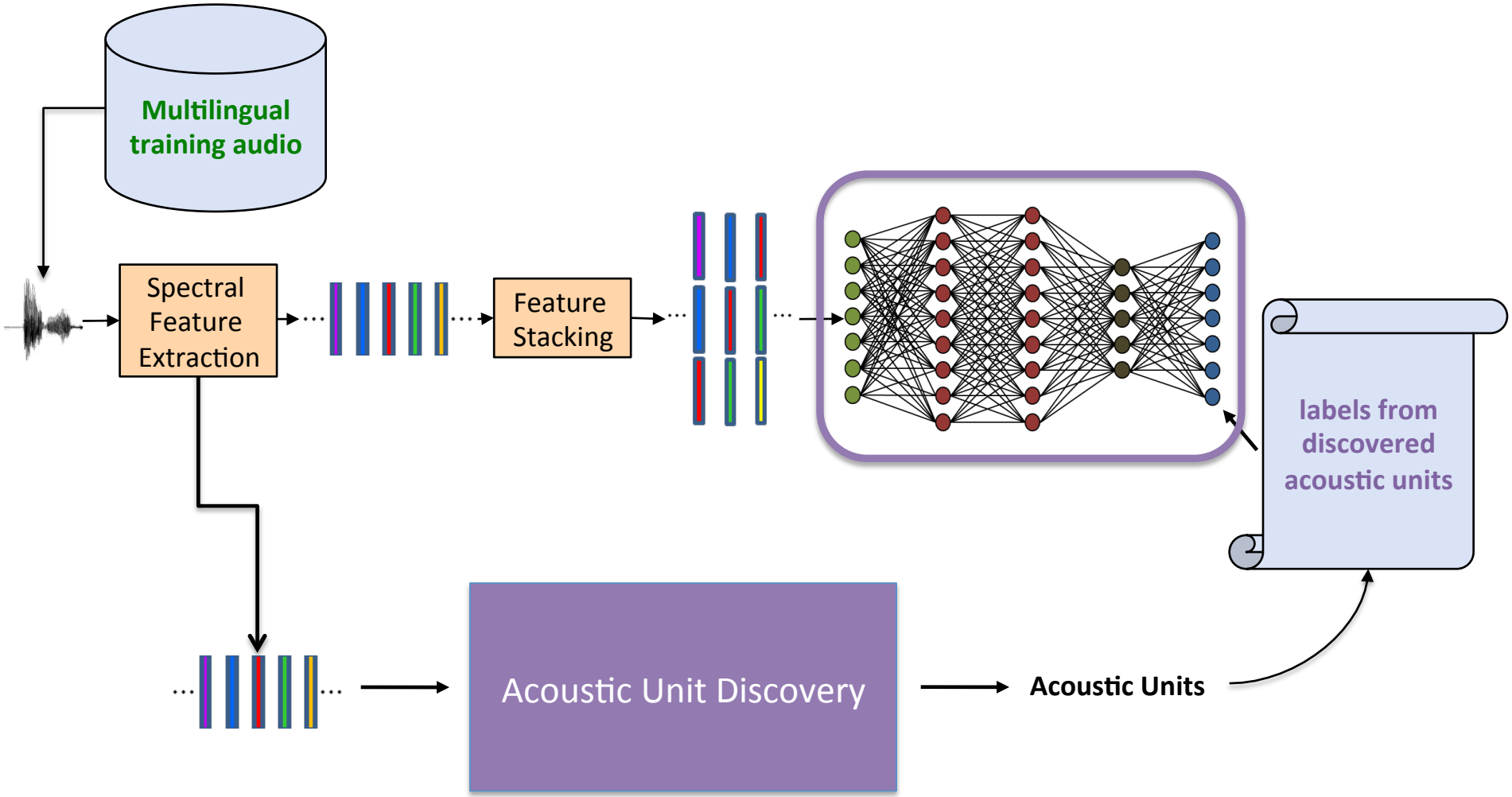


S. Shum, D. Harwath, N. Dehak, and J. Glass, "On the Use of Acoustic Unit Discovery for Language Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, 2016. [To Appear]

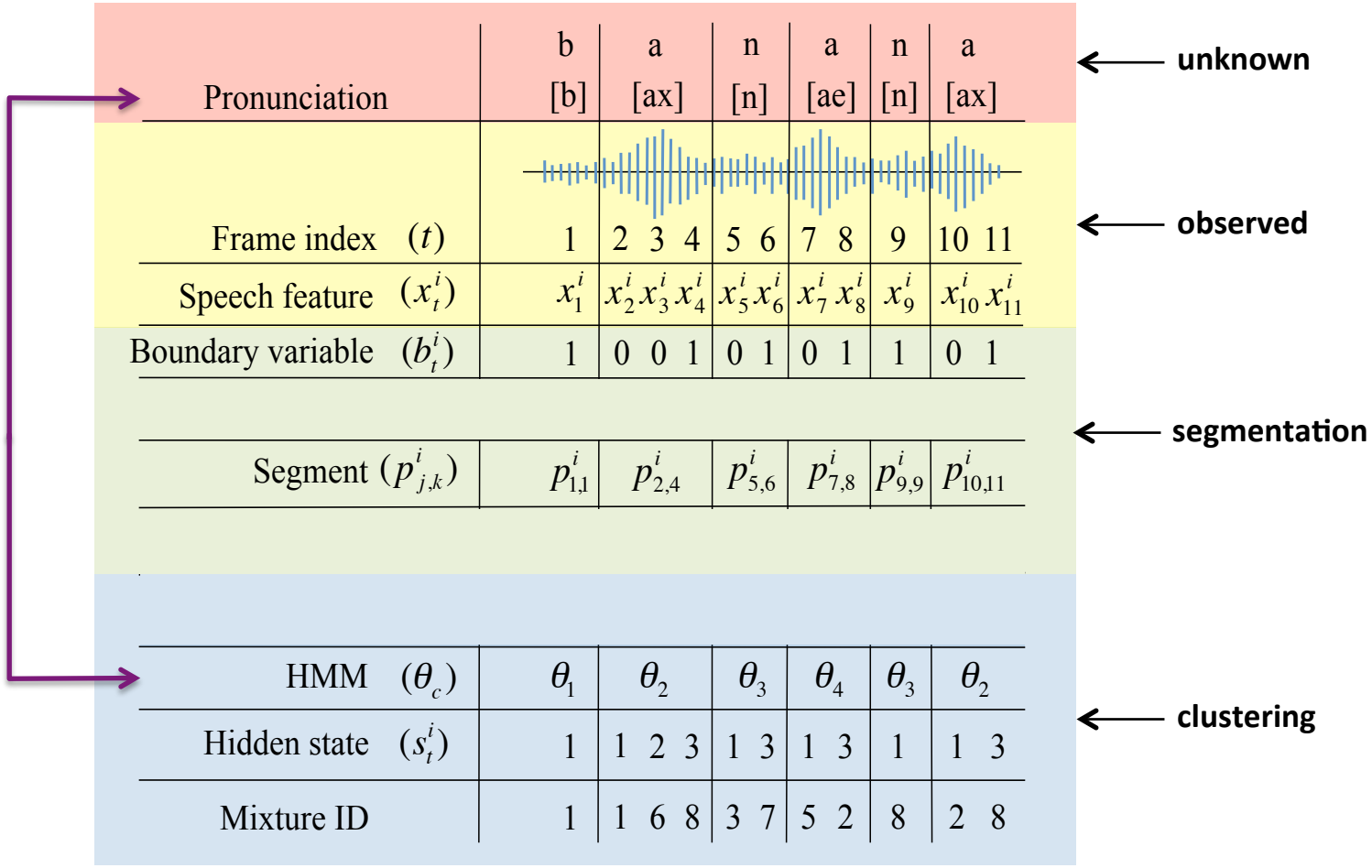
Motivation

- Standard speech recognition systems rely on
  - Transcribed speech
  - Language models
  - Pronunciation dictionaries
- Usually only available for a subset of languages
- Can we discover what we need automatically?
- Unsupervised methods allow us to work directly on the (untranscribed) data pertaining to the evaluation at hand

Assessing the usefulness of acoustic unit discovery for language ID



Bayesian acoustic unit discovery (BAUD)



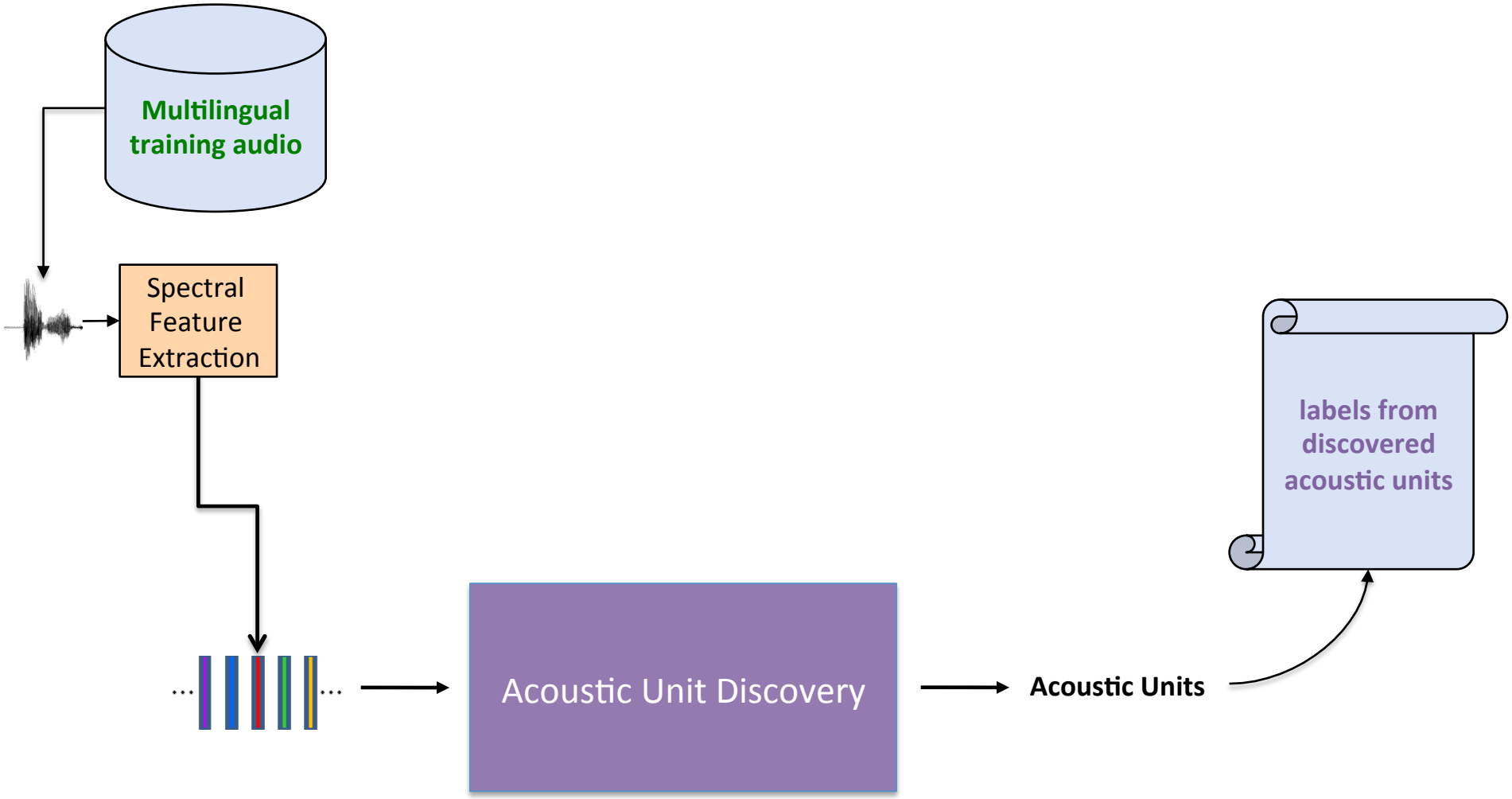
Chia-ying Lee and James R. Glass, "A Nonparametric Bayesian Approach to Acoustic Model Discovery," *Proceedings of ACL*, 2012.

Approximate Distributed BAUD

- Based off of work in (Lee & Glass, 2012)
- Not quite fully Bayesian
  - Specify number of acoustic units to learn (100)
  - Parallelization only *approximates* Gibbs sampling
    - Serial Gibbs sampling takes much longer to converge
    - But scalable to larger datasets (200+ hours) than TIMIT
  - Maximum likelihood model updates

A. Ihler and D. Newman, "Understanding Errors in Approximate Distributed Latent Dirichlet Allocation," *IEEE Transactions on Knowledge and Data Engineering*, May 2012.

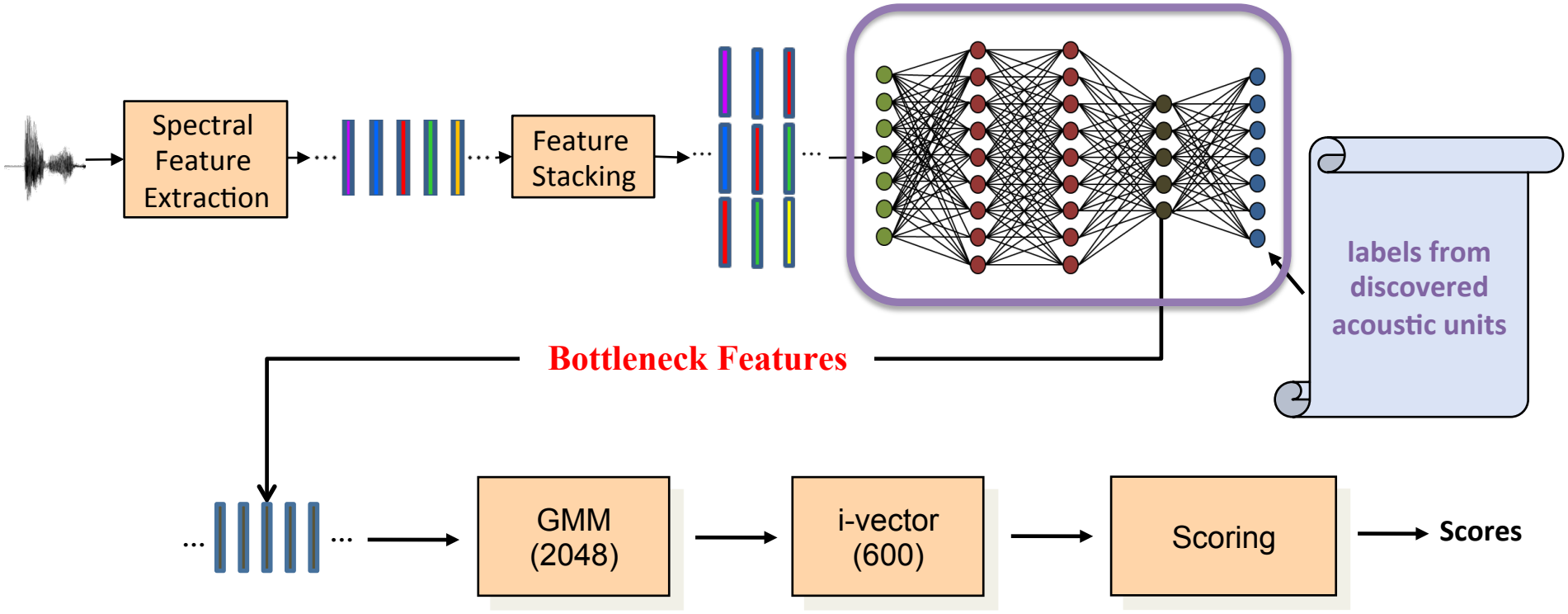
Assessing the usefulness of acoustic unit discovery for language ID



Roadmap

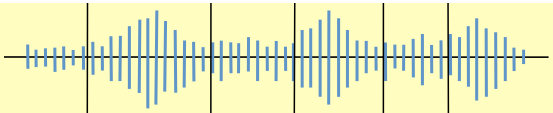
- Language recognition using i-vectors
  - Spectral feature baseline
  - DNN bottleneck feature benchmark
- Parallelizing a Bayesian nonparametric model for large-scale acoustic unit discovery
- Experiments
  - The usefulness of context-dependent modeling
  - The magic of fusion
  - The impact of improved acoustic features
  - The generalizability of language-specific perspectives

Acoustic unit discovery-based bottleneck i-vector system (proposed)



S. Shum, D. Harwath, N. Dehak, and J. Glass, "On the Use of Acoustic Unit Discovery for Language Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, 2016. [To Appear]

Per-frame label sequences



Frame index (t)	1	2	3	4	5	6	7	8	9	10	11
Speech feature (x_t^i)	x_1^i	x_2^i	x_3^i	x_4^i	x_5^i	x_6^i	x_7^i	x_8^i	x_9^i	x_{10}^i	x_{11}^i
Boundary variable (b_t^i)	1	0	0	1	0	1	0	1	1	0	1
Segment $(p_{j,k}^i)$	$p_{1,1}^i$	$p_{2,4}^i$			$p_{5,6}^i$		$p_{7,8}^i$		$p_{9,9}^i$	$p_{10,11}^i$	
HMM (θ_c)	θ_1	θ_2			θ_3		θ_4		θ_3	θ_2	
Hidden state (s_t^i)	1	1	2	3	1	3	1	3	1	1	3
Mixture ID	1	1	6	8	3	7	5	2	8	2	8

← unit sequence

← state sequence

Chia-ying Lee and James R. Glass, "A Nonparametric Bayesian Approach to Acoustic Model Discovery," *Proceedings of ACL*, 2012.

Exploiting context-dependence

- Treat unit sequences as transcriptions and train a unit recognizer
  - Relaxes boundary variable-based segmentation
  - Allows for context-dependent modeling of units

- Use resulting context-dependent HMM state sequences (i.e., “senones”) as per-frame labels for DNN training

Initial experiments and results

- Run BAUD on 240hrs of multilingual audio
  - 10 hours from each of 24 languages represented

	100 units (CI)	300 states (CI)	4000 senones (CD)
Multilingual BAUD (240 hrs)	9.0	6.7	**5.2**
Spectral Feature Baseline		5.3	
Transcribed English Benchmark		**2.6**	

Results shown are the Average Detection Cost * 100 and can be interpreted as a measure of error.

Finding complementarity

- BAUD system was barely better than baseline
  - But what if we fused the two systems together?

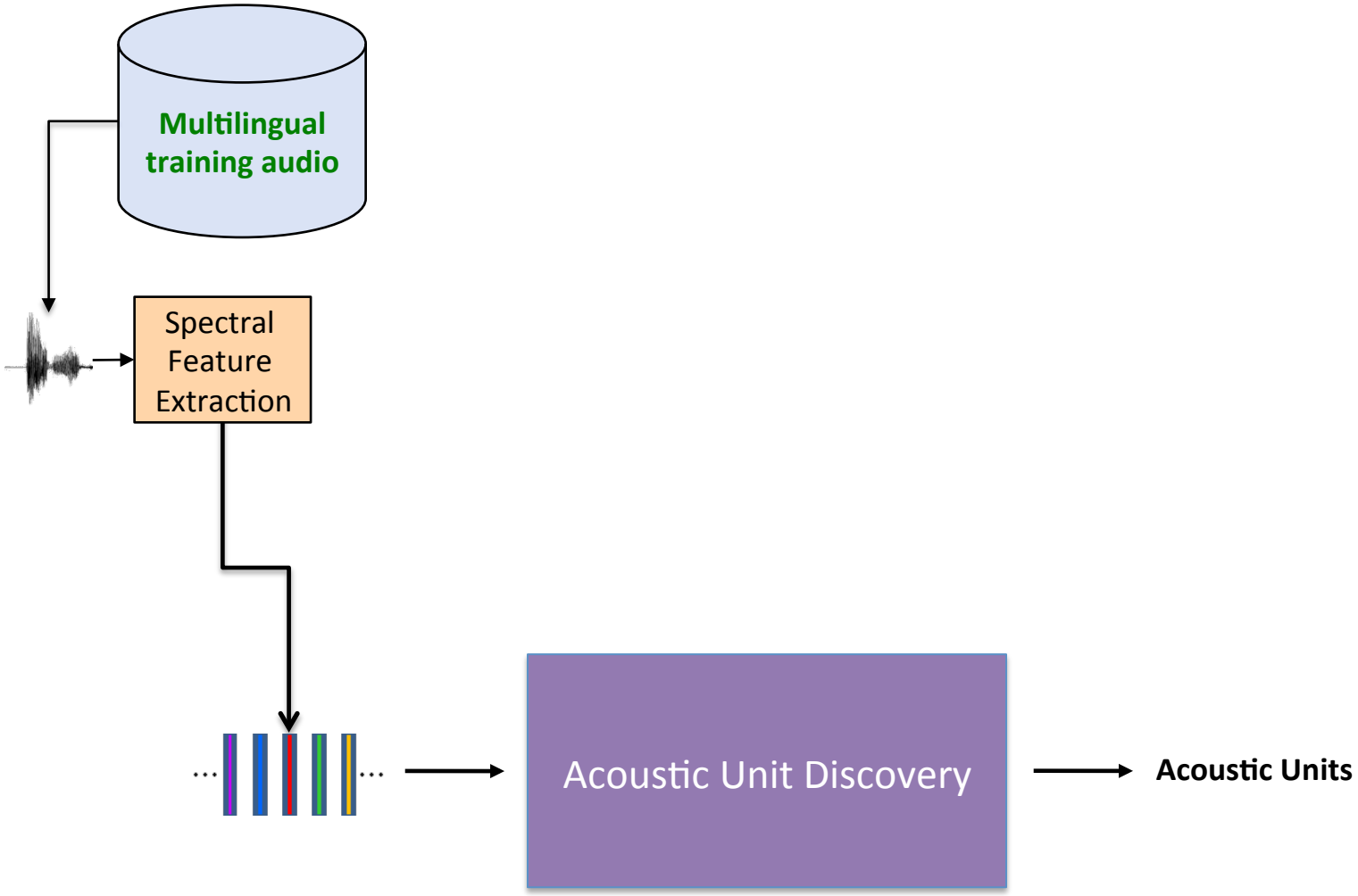
	Detection Cost (30s Segments)
[*] Spectral Feature Baseline	5.3
[*] BAUD(LRE), senones	5.2
Score-level fusion of [*] above	**3.8**
Transcribed SWB Benchmark	2.6

Can we do better with improved features?

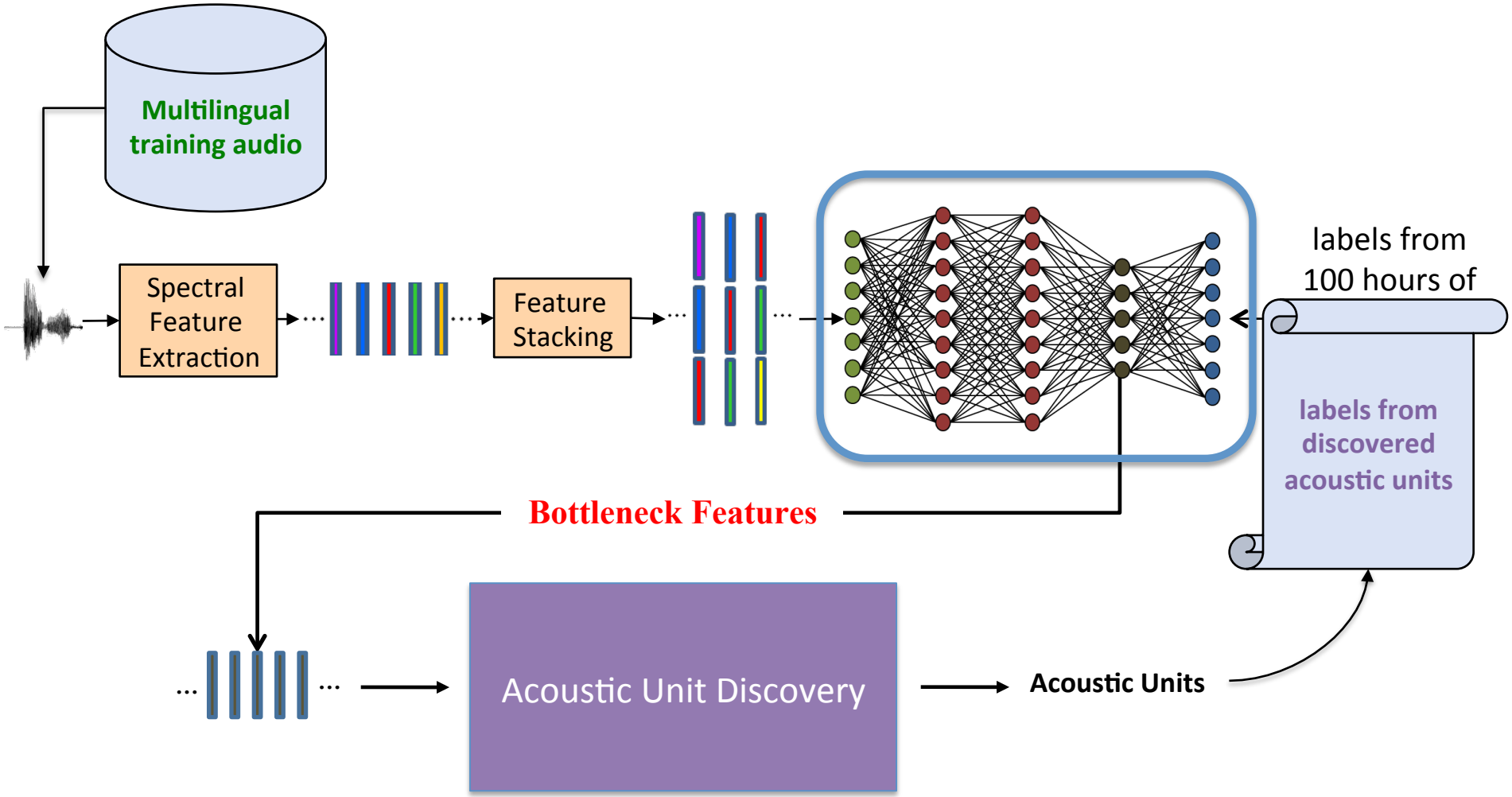
- Unsupervised methods make assumptions about the distribution of the data
  - Is there a representation that better fits these assumptions?

- Experiment
  - Run BAUD using bottleneck features trained on 100 hours of transcribed English
  - No longer fully unsupervised, but neither unreasonable nor unrealistic

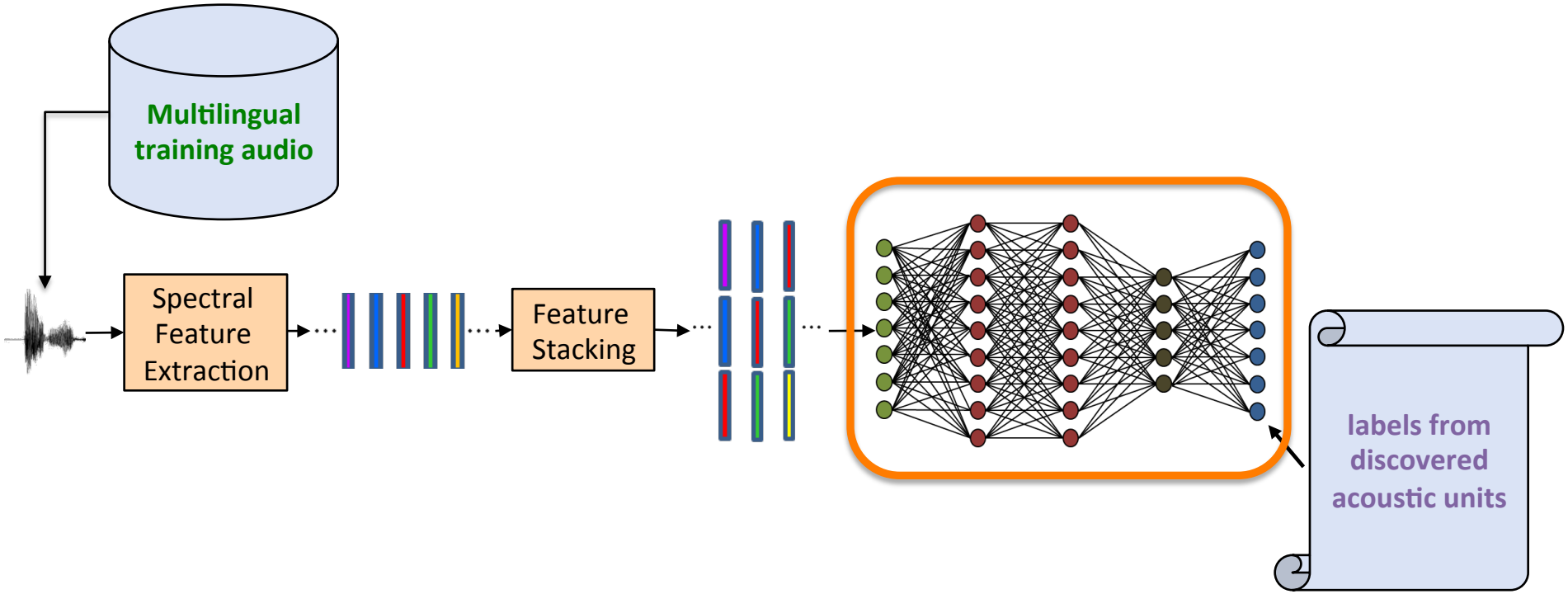
Vanilla BAUD



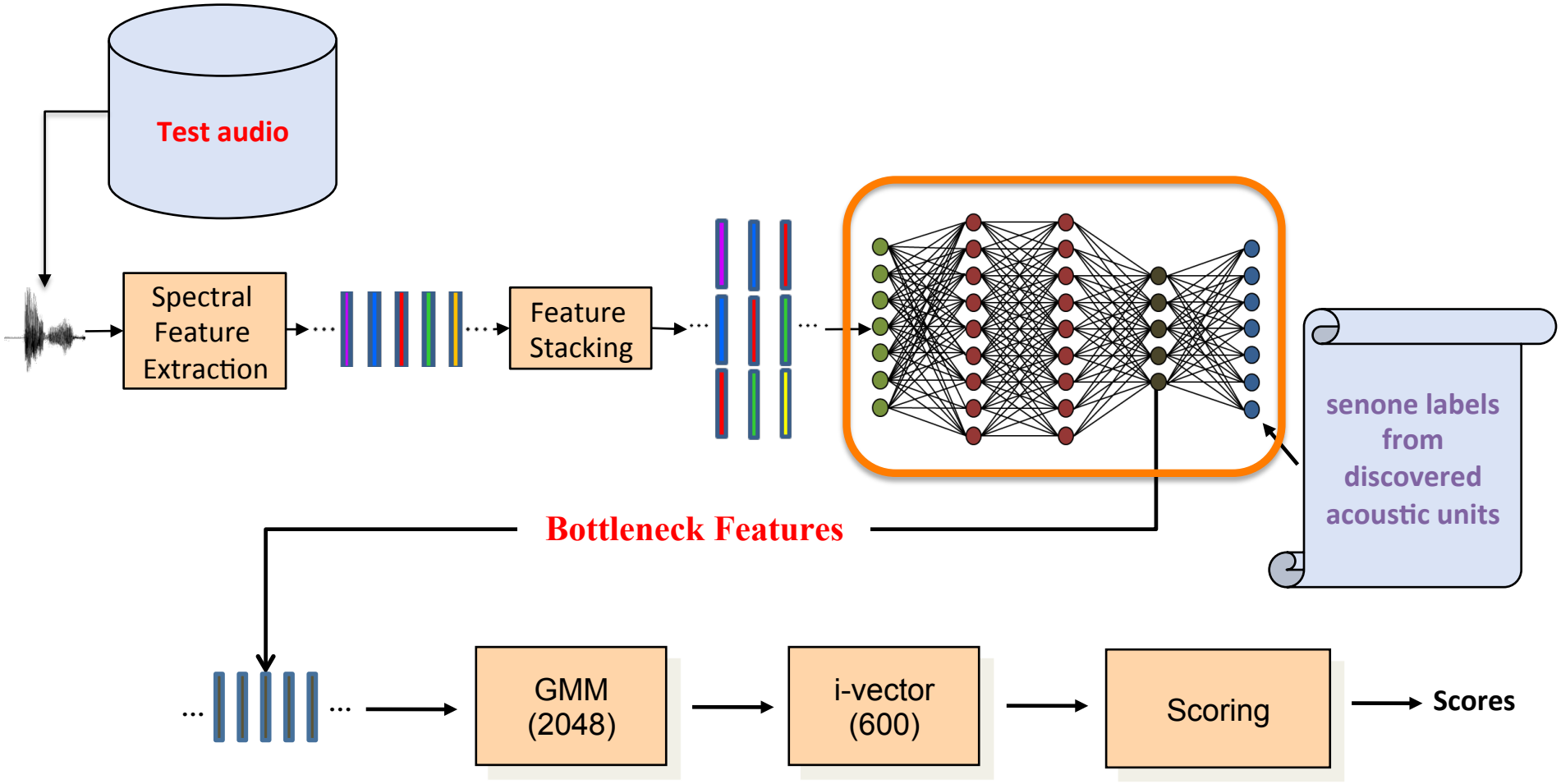
Transcribed English-based bottleneck features for BAUD



Train a new DNN from scratch



English-inspired acoustic unit discovery-based bottleneck i-vector system



Transcribed English-inspired improvements

	Detection Cost (30s Segments)
[*] Spectral Feature Baseline	5.3
[*] BAUD(LRE, MFCC), senones	5.2
Score-level fusion of [*] above	3.8
[**] BAUD(LRE, SWB-BN), senones	**2.9**
[**] Transcribed SWB Benchmark	2.6
Score-level fusion of [**] above	**2.1**

Roadmap

- Language recognition using i-vectors
  - Spectral feature baseline
  - DNN bottleneck feature benchmark
- Parallelizing a Bayesian nonparametric model for large-scale acoustic unit discovery
- Experiments
  - The usefulness of context-dependent modeling
  - The magic of fusion
  - The impact of improved acoustic features
  - The generalizability of language-specific perspectives

Roadmap

- Language recognition using i-vectors
  - Spectral feature baseline
  - DNN bottleneck feature benchmark
- Parallelizing a Bayesian nonparametric model for large-scale acoustic unit discovery
- Experiments
  - The usefulness of context-dependent modeling
  - The magic of fusion
  - The impact of improved acoustic features
  - The generalizability of language-specific perspectives

NIST Language Recognition Evaluation 2015

- 20 languages, 6 clusters
  - Arabic
    - Egyptian, Iraqi, Levantine, Maghrebi, Modern Standard
  - Chinese
    - Cantonese, Mandarin, Min, Wu
  - English
    - British, American, Indian
  - French
    - West African, Haitian Creole
  - Iberian
    - Caribbean Spanish, European Spanish, Latin American Spanish, Brazilian Portuguese
  - Slavic
    - Polish, Russian

NIST Language Recognition Evaluation 2015

- Different evaluation protocol
  - Identify individual languages within their respective language clusters
- Same performance trends
- Language-specific perspectives

NIST Language Recognition Evaluation 2015

- 20 languages, 6 clusters
  - Arabic
    - Egyptian, Iraqi, Levantine, Maghrebi, Modern Standard
  - Chinese
    - Cantonese, Mandarin, Min, Wu
  - English
    - British, American, Indian
  - French
    - West African, Haitian Creole
  - Iberian
    - Caribbean Spanish, European Spanish, Latin American Spanish, Brazilian Portuguese
  - Slavic
    - Polish, Russian

NIST Language Recognition Evaluation 2015

- 20 languages, 6 clusters
  - Arabic
    - Egyptian, Iraqi, Levantine, Maghrebi, Modern Standard
  - Chinese
    - Cantonese, Mandarin, Min, Wu
  - English
    - British, American, Indian
  - French
    - West African, Haitian Creole
  - Iberian
    - Caribbean Spanish, European Spanish, Latin American Spanish, Brazilian Portuguese
  - Slavic
    - Polish, Russian

NIST LRE 2015
(Vanilla BAUD)

Cluster	# hrs	Arabic	Chinese	English	Iberian	Slavic	Average
Arabic	23	**25.8**	21.4	18.6	22.9	8.84	19.5
Chinese	23	26.7	**21.0**	17.5	22.9	10.2	19.6
English	23	26.5	21.8	**15.7**	22.0	9.37	19.3
Iberian	23	27.3	21.5	19.8	**22.5**	9.71	20.2
Slavic	23	26.1	21.2	19.1	22.5	**8.69**	19.5
Fused	(115)	24.9	18.4	14.2	20.7	7.00	**17.0**
All	115	25.3	18.2	16.4	22.0	7.89	18.0

NIST LRE 2015
(English-inspired BAUD)

Cluster	# hrs	Arabic	Chinese	English	Iberian	Slavic	Average
Arabic	23	**20.9**	16.0	15.2	20.3	6.39	15.8
Chinese	23	22.1	16.1	15.2	20.3	5.72	15.9
English	23	21.6	15.4	**12.8**	19.2	5.84	15.0
Iberian	23	21.4	**15.3**	15.5	**19.1**	5.40	15.3
Slavic	23	21.3	16.0	15.6	20.8	**4.66**	15.7
Fused	(115)	19.5	12.9	11.2	17.6	3.53	**12.9**
Benchmark	315	19.6	13.1	11.2	18.4	3.27	13.1

Summary

- Language recognition using i-vectors
  - Spectral feature baseline
  - DNN bottleneck feature benchmark
- Parallelizing a Bayesian nonparametric model for large-scale acoustic unit discovery
- Experiments
  - The usefulness of context-dependent modeling
  - The magic of fusion
  - The impact of improved acoustic features
  - The generalizability of language-specific perspectives

Overview

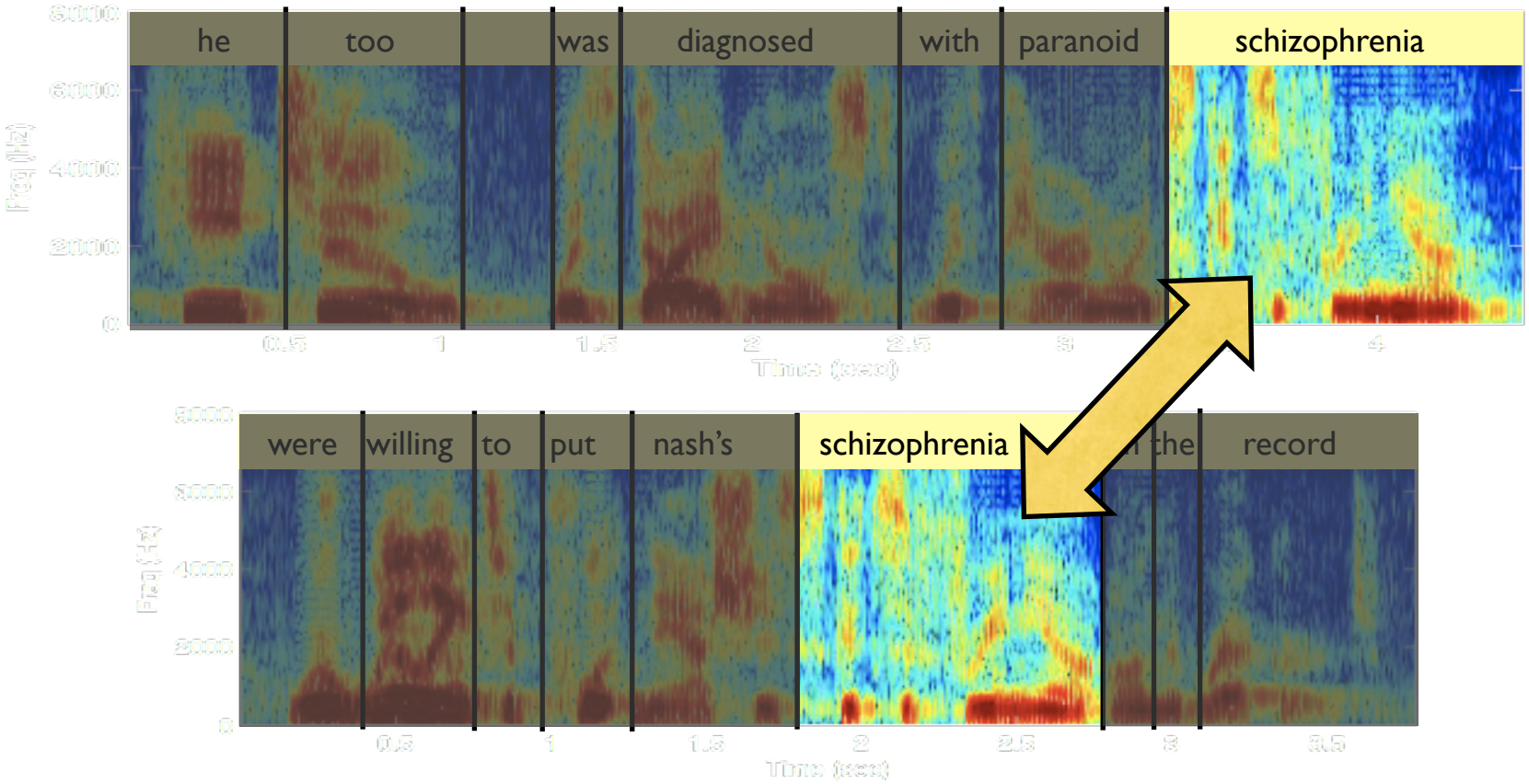
	Domain Adaptation	Weak Supervision
Speaker Verification	Adapt system to changes in recording technology by applying existing models to new, unlabeled data sets	Actively explore a database of unknown speakers and build speaker models using pairwise equivalence constraints
Language Identification	Augment existing volumes of transcribed speech with large-scale, unsupervised discovery of acoustic units on untranscribed, multilingual data	Use equivalence constraints between acoustic sequences to improve speaker-independence of discovered acoustic units

Overview

	Domain Adaptation	Weak Supervision
Speaker Verification	Adapt system to changes in recording technology by applying existing models to new, unlabeled data sets	Actively explore a database of unknown speakers and build speaker models using pairwise equivalence constraints
Language Identification	Augment existing volumes of transcribed speech with large-scale, unsupervised discovery of acoustic units on untranscribed, multilingual data	Use equivalence constraints between acoustic sequences to improve speaker-independence of discovered acoustic units

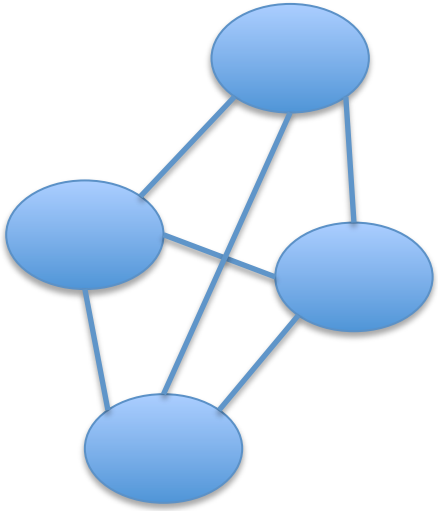
Can we improve acoustic unit discovery using equivalence constraints?

- Find repeated acoustic segments



Can we improve acoustic unit discovery using equivalence constraints?

- Find repeated acoustic segments



- Verify that these segments match

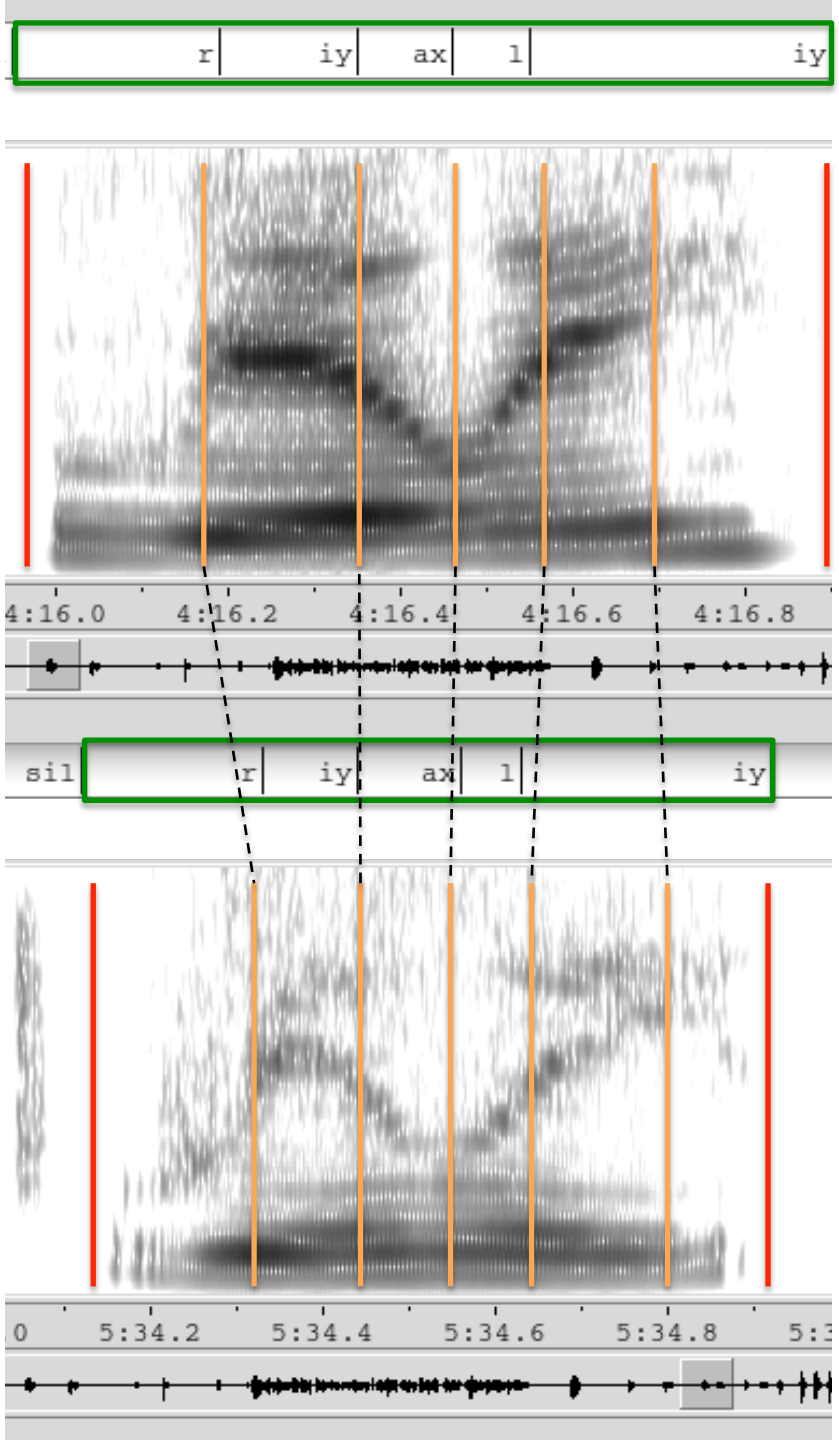
- Constrain unit discovery process to learn similar unit sequences for matched segments

Proposed Methods

- Assumption
  - Given sets of pronunciation-equivalent utterances (e.g., words, phrases, or sentences)

- DTW-based segmentation consolidation
- Equivalence-constrained clustering

Transcription:
“really”
(not known)



DTW-based segmentation consolidation

- For each set of pronunciation-equivalent utterances:

- Pick an utterance to use as an exemplar
- Obtain landmark segmentation from exemplar

- Use dynamic time warp (DTW) alignment between exemplar and all other utterances to map exemplar segmentation to all the other utterances

Equivalence-constrained clustering

- For each set of pronunciation-equivalent utterances:

- Pick an utterance uniformly at random;
- Sample acoustic unit sequence (as in BAUD);

- Pretend as though *every other utterance* in the set also sampled the *exact same acoustic unit sequence* and update models accordingly.

Key Takeaways

- Experiments on TIMIT
  - Run constrained BAUD on training subset
  - Evaluate models on test subset
- Evaluation metrics and results
  - Normalized mutual information (NMI)
    - ~5% relative increase (vs. unconstrained BAUD)
  - Defined a word error rate-based metric to measure inconsistency between equivalent sequences
    - ~10% relative decrease (vs. unconstrained BAUD)

Thesis Contributions

	Domain Adaptation	Weak Supervision
Speaker Verification	Adapt system to changes in recording technology by applying existing models to new, unlabeled data sets	Actively explore a database of unknown speakers and build speaker models using pairwise equivalence constraints
Language Identification	Augment existing volumes of transcribed speech with large-scale, unsupervised discovery of acoustic units on untranscribed, multilingual data	Use equivalence constraints between acoustic sequences to improve speaker-independence of discovered acoustic units

Future Work

- Domain adaptation
  - Telephone → Microphone
  - Out-of-domain detection
- Weak supervision
  - Noisy labels
  - Improved feature representations
  - Towards crowd-supervised development of speech technologies

Acknowledgments