CSAIL

MIT COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE LABORATORY

From Vectors Representing Speech to Graphs Representing Corpora

Stephen Shum

**With Najim Dehak, Jim Glass, Doug Reynolds, Bill Campbell, and many others*

November 2013

CSAIL

MIT COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE LABORATORY

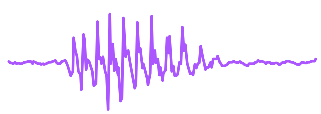
From Vectors Representing Speech to Graphs Representing Corpora:

Reconciling how far we've come with how far we still have to go

Extracting Information from Speech

Goal: Automatically extract information transmitted in speech signal

Speech Signal



Speech Recognition → Words
"How are you?"

Language Recognition → Language Name
English

Speaker Recognition → Speaker Name
James Wilson

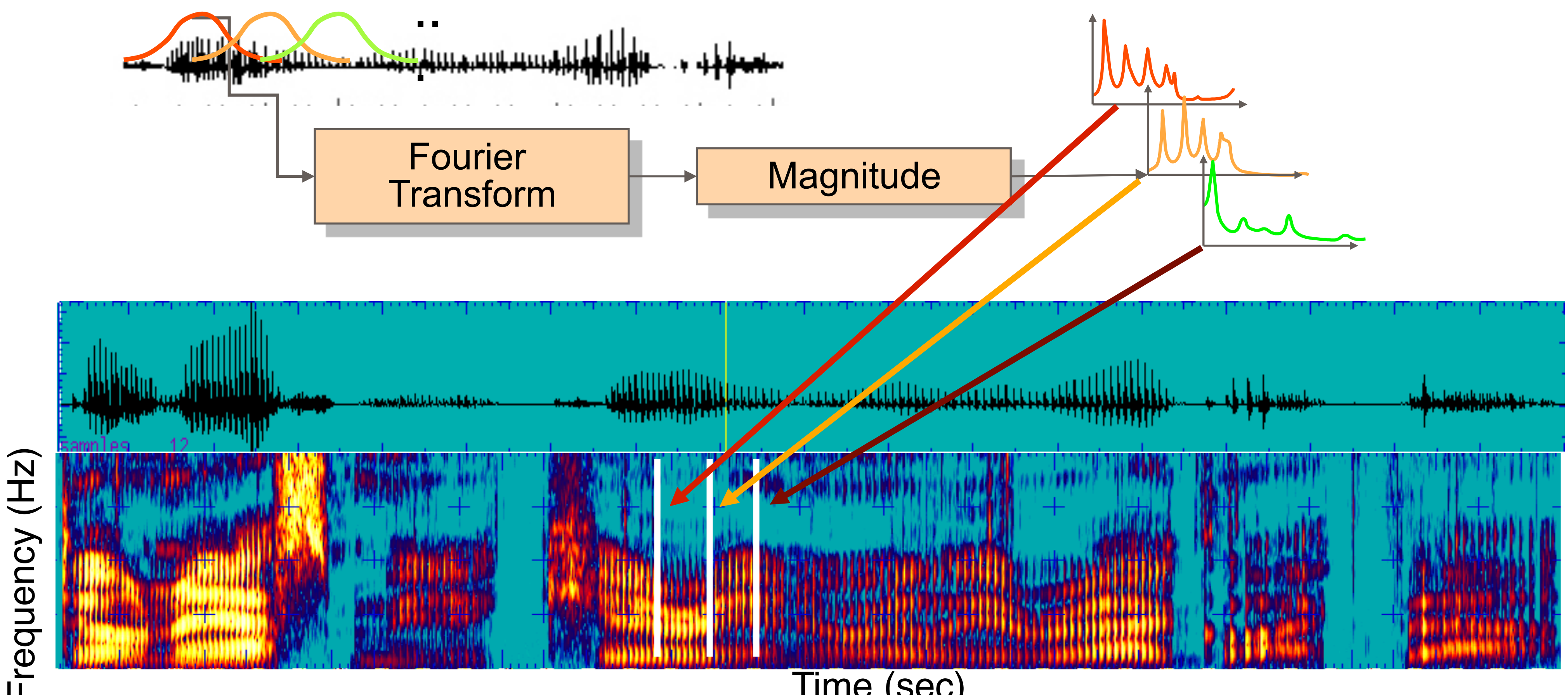
Emotion Recognition → Emotion category
Happy, angry …

Roadmap

- Vector-based representations of speech
- Graph-based representation of audio databases
- Domain adaptation for speaker recognition

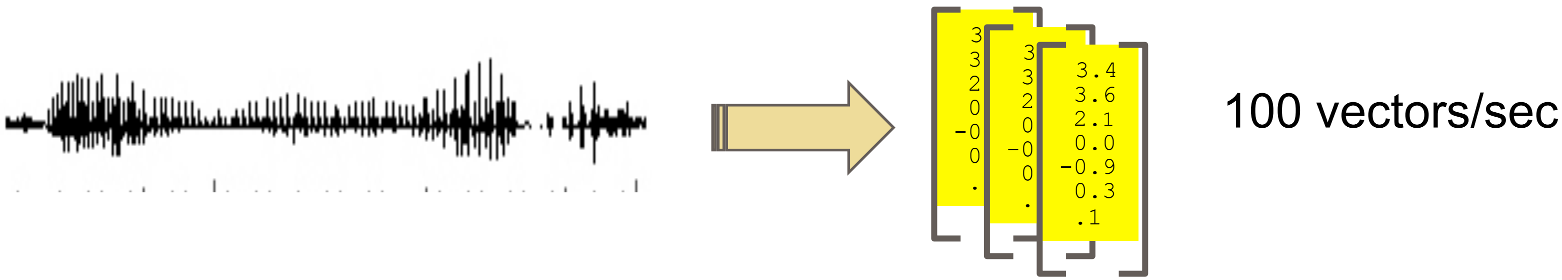
Information in Speech

- **Speech is a time-varying signal whose information can be observed in the time and frequency domains**
  – Such information can be captured via a time sequence of features

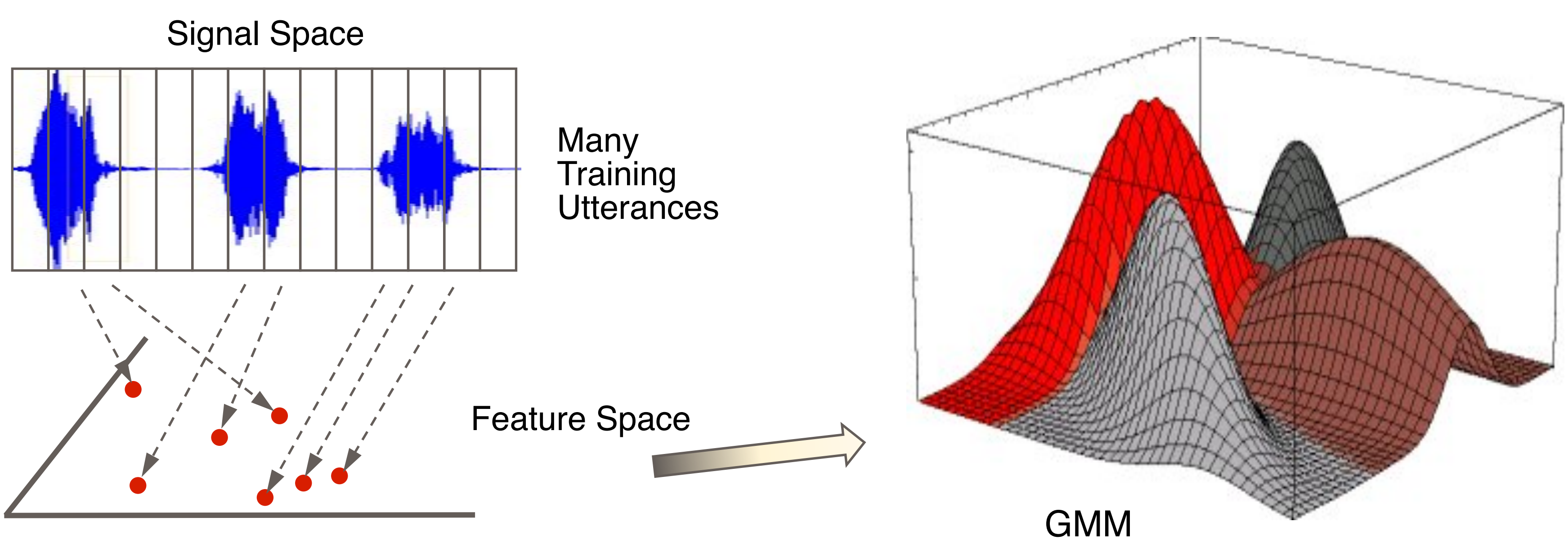


Modeling Feature Sequences with GMMs

- **We need to model the distribution of feature vector sequences**
  - e.g., Mel Frequency Cepstral Coefficients (MFCCs)

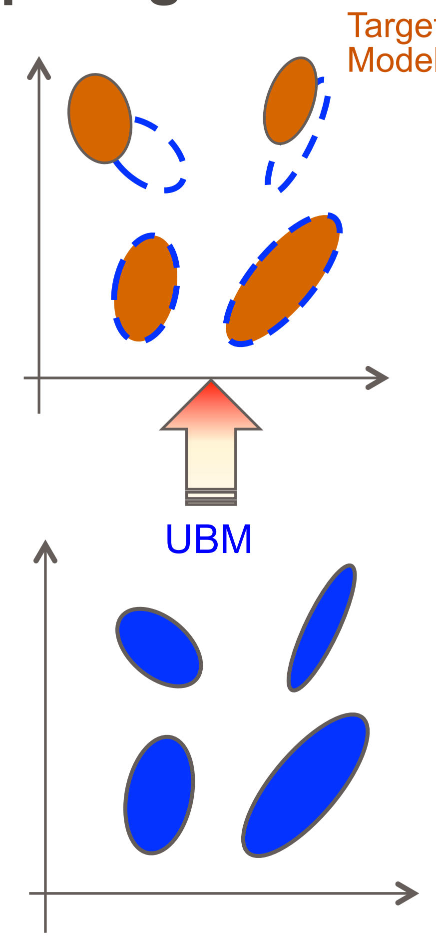


- **Gaussian mixture models (GMMs) are a common representation**

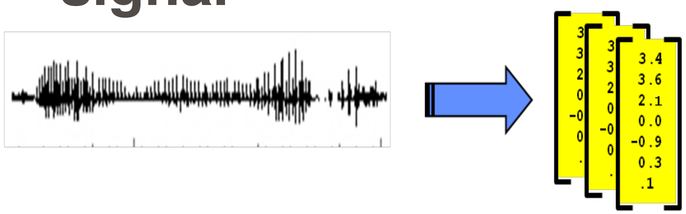


Modeling with Adapted GMM-UBMs

(3) Adapt target model from UBM



(2) Extract feature vector sequence from speech signal



(1) Start with UBM

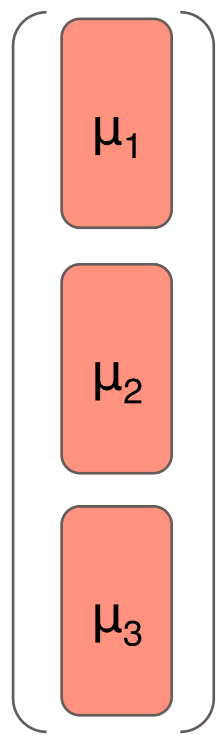
GMM-UBM and MAP Adaptation

- **Target model is trained by adapting from background model**
  - Couples models together and helps with limited target training data
- **Adaptation only updates mean parameters representing acoustic events seen in target training data**
  - Sparse regions of feature space filled in by UBM mean parameters
    - **Both an advantage and a disadvantage**
- **Disadvantage**
  - Limited target training data still prevents some UBM components from being adapted.

Advantages

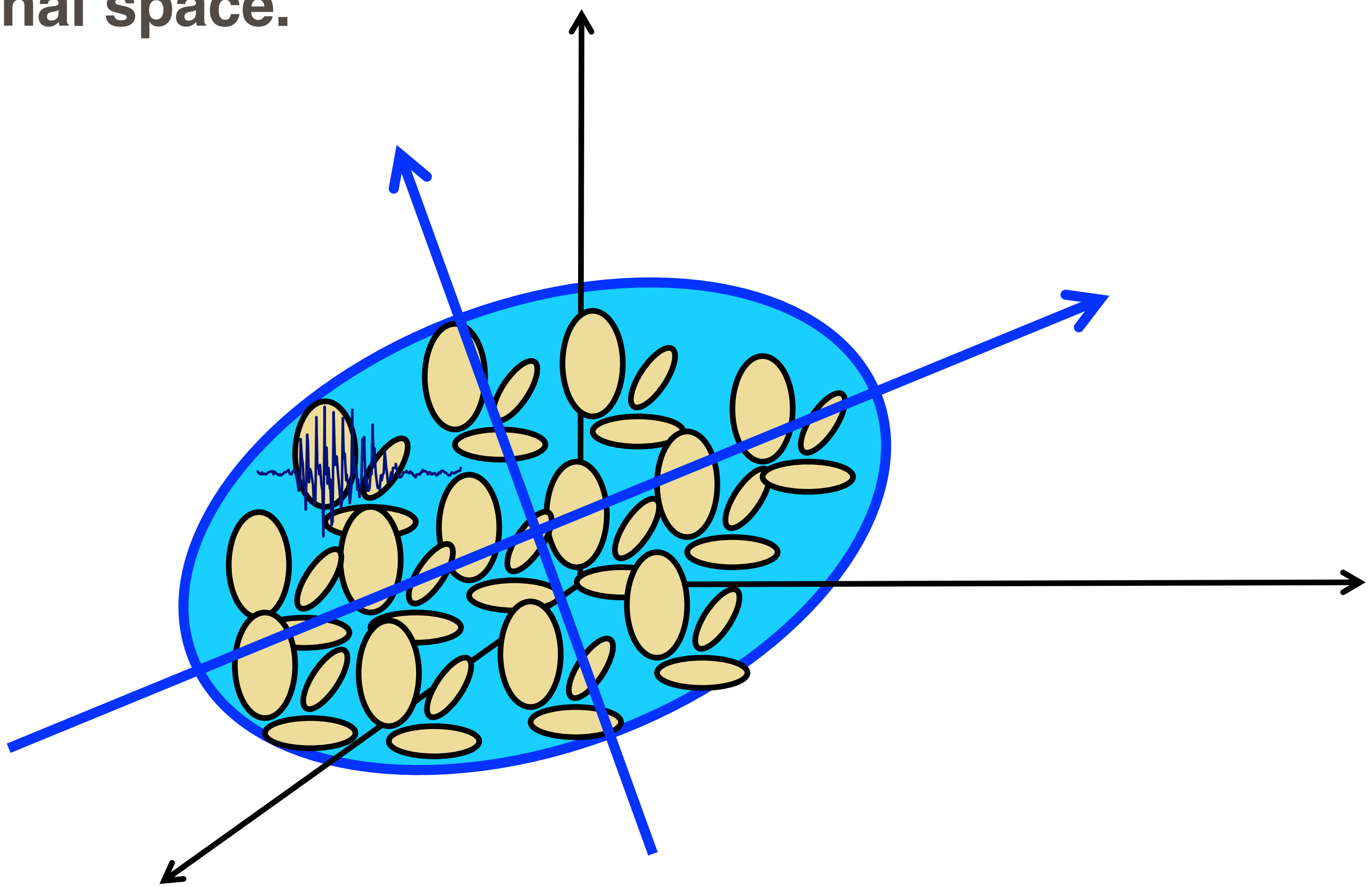
- **Re-parameterize GMM as a *supervector*.**
  - Concatenate all mixture mean components of a GMM.

- **The way the UBM adapts to a given speaker ought to be somewhat constrained.**
  - Regardless of speaker identity, there should exist at least some correspondence in the way the means move relative to one another.



The Total Variability Space

- Suppose a GMM supervector corresponds to a point in high-dimensional space.

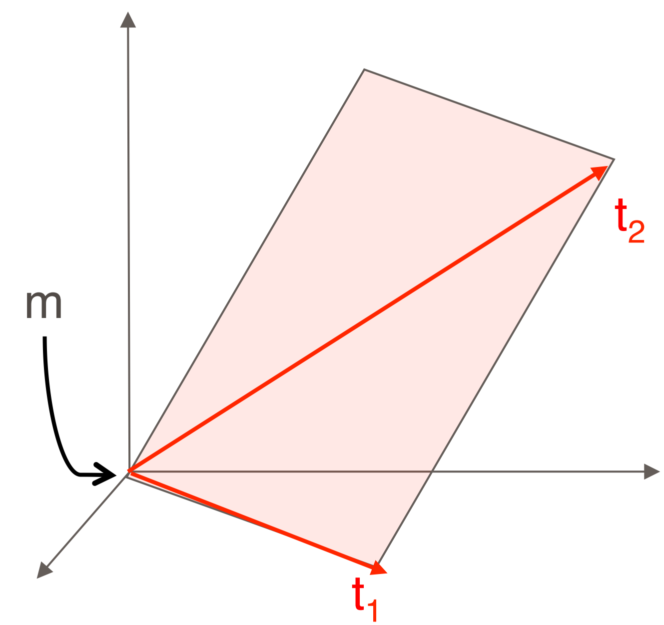


- Use factor analysis to capture the directions of maximum between-utterance variability.

The Total Variability Approach

- **Assumption (Dehak, 2009)**
  - All pertinent variabilities lie in some low dimensional subspace T
    - **Call it the Total Variability Space**

$$\mathbf{M} = \mathbf{m} + \mathbf{T}\mathbf{w}$$



* **$\mathbf{w}$ is the i-vector (identity/intermediate vector)**

Regarding i-vectors

- **For some speech segment s, its associated i-vector w_s can be seen as a low-dimensional summary of that segment's distribution of acoustic features (with respect to a UBM).**

- **(Relatively) low-dimensional random vector (600 << 120,000)**
  - Standard normal prior distribution, *N(0, I)*

- **Given some speech,**
  - Posterior mean → i-vector
  - Posterior covariance → i-vector covariance

Recap

- **Model variable-length sequences of acoustic features using a GMM adapted from a UBM.**

- **Re-parameterize the GMM into a high-dimensional *supervector* by concatenating all mixture means.**

- **Obtain a lower-dimensional *i-vector* representation via factor analysis, which uses a Total Variability subspace to model directions of maximal variability in the supervector space.**

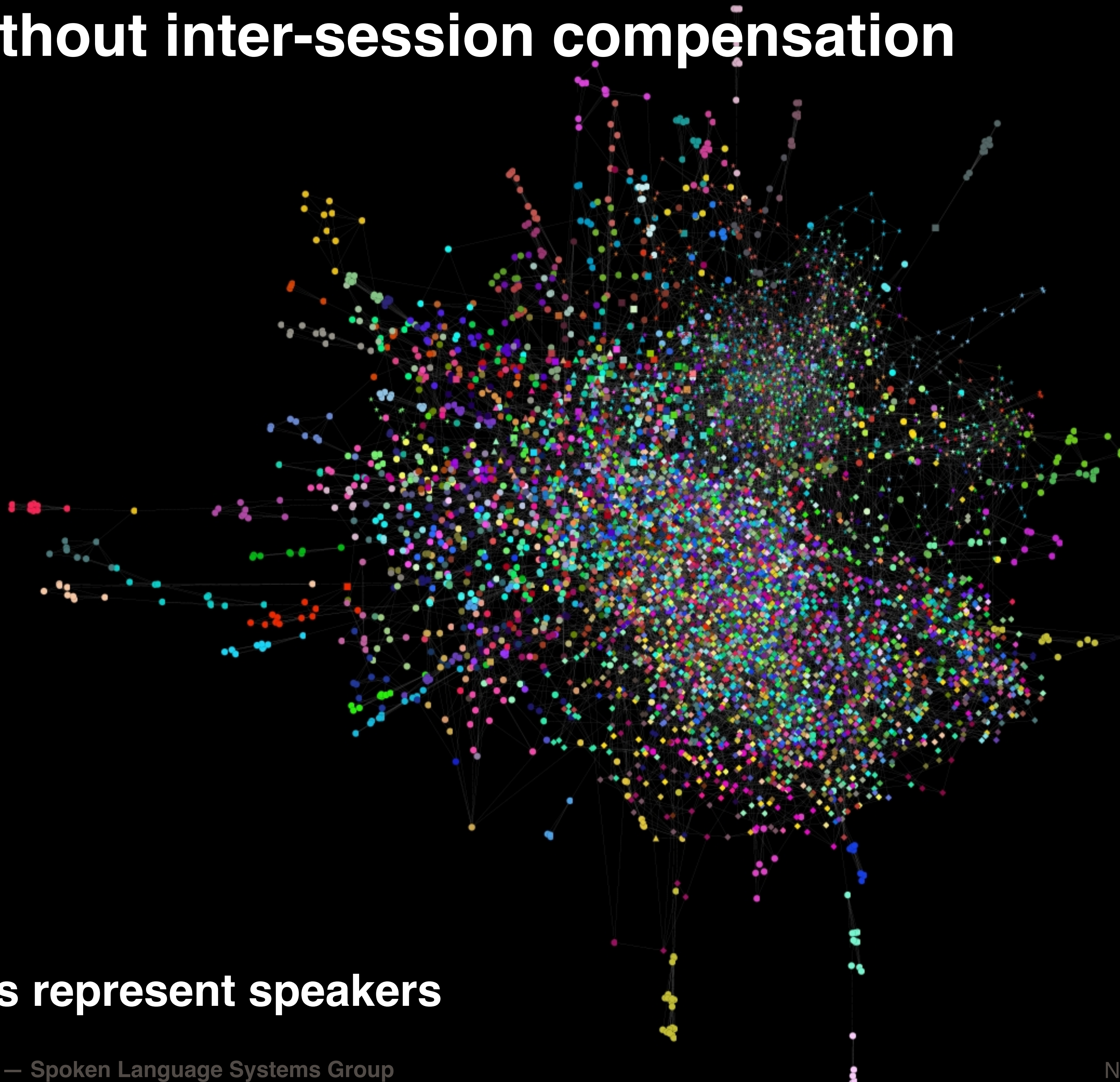
Exploiting the convenience of a vector-based representation

- **Allows for rote application of machine learning techniques to compensate for unwanted channel/inter-session variabilities**
  - Nuisance Attribute Projection (NAP)
  - Linear Discriminant Analysis (LDA) + Within-Class Covariance Normalization (WCCN) + cosine scoring
  - Probabilistic LDA (PLDA)

Effects of inter-session compensation

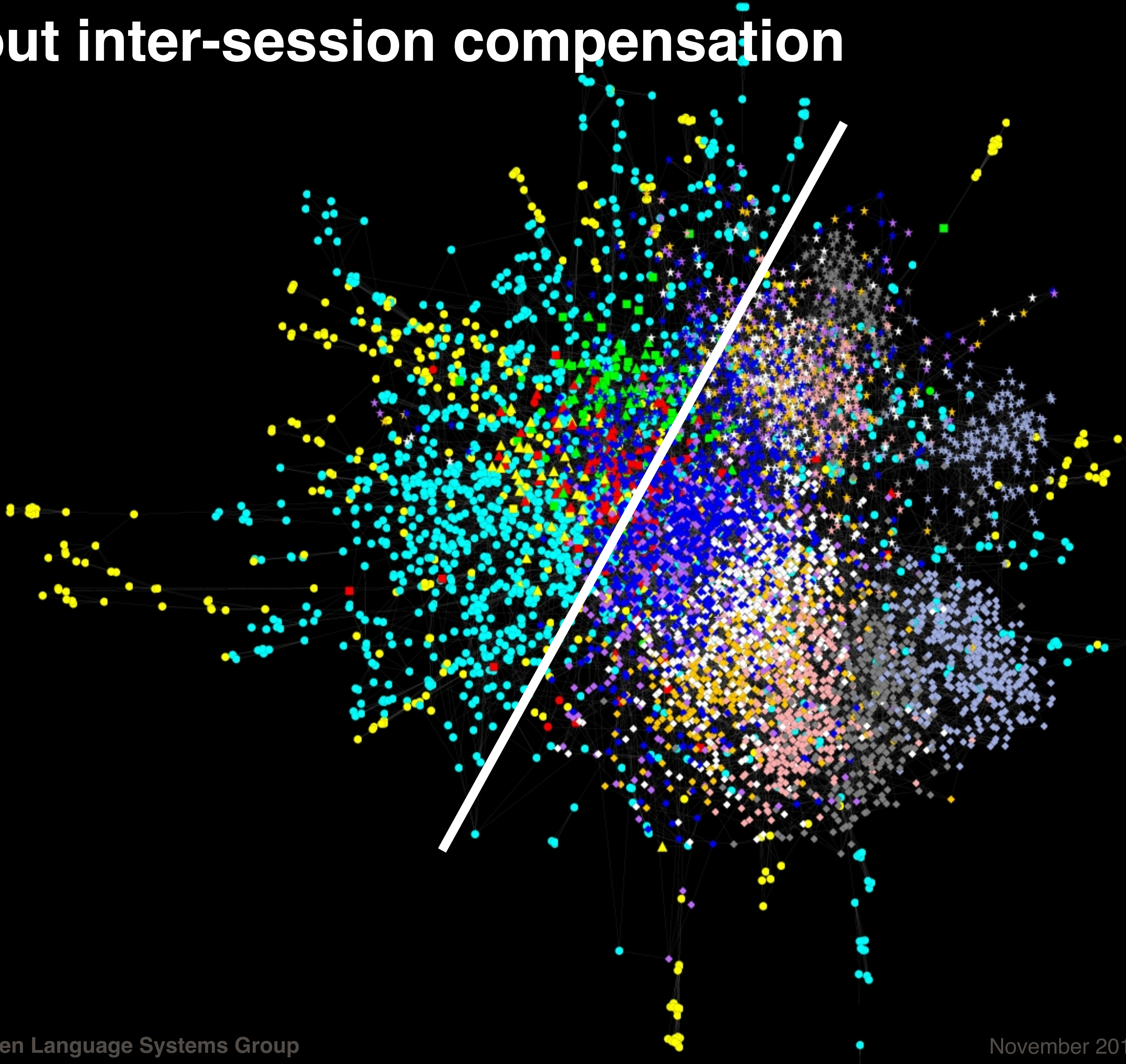
- **Graph visualization**
  - Represent each segment as a node with connections (edges) to its K nearest neighbors (K-NN); K = 3
  - Absolute locations of the nodes are not important
  - Relative locations of nodes provide information about connectedness and similarity

Without inter-session compensation



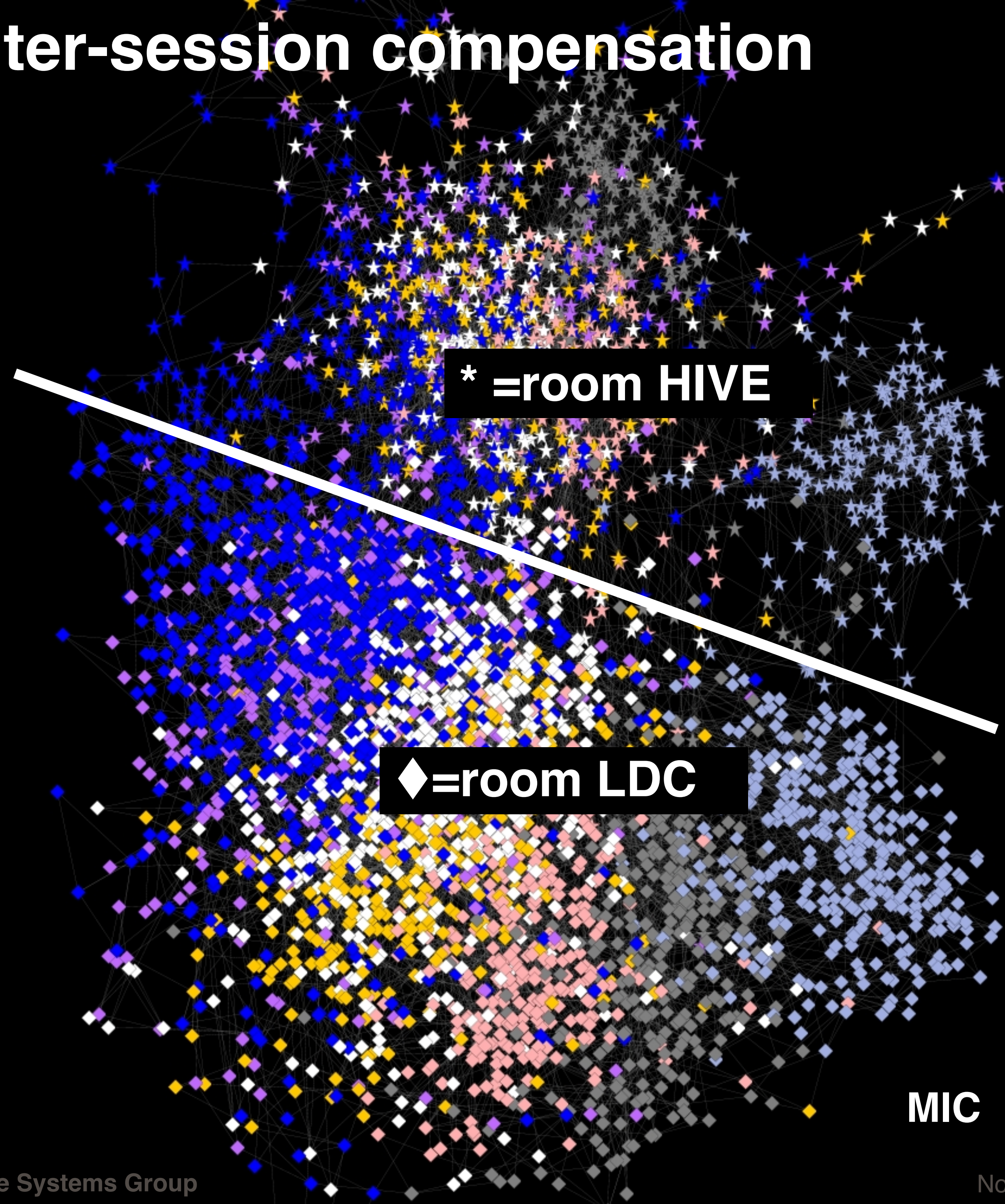
Colors represent speakers

Without inter-session compensation



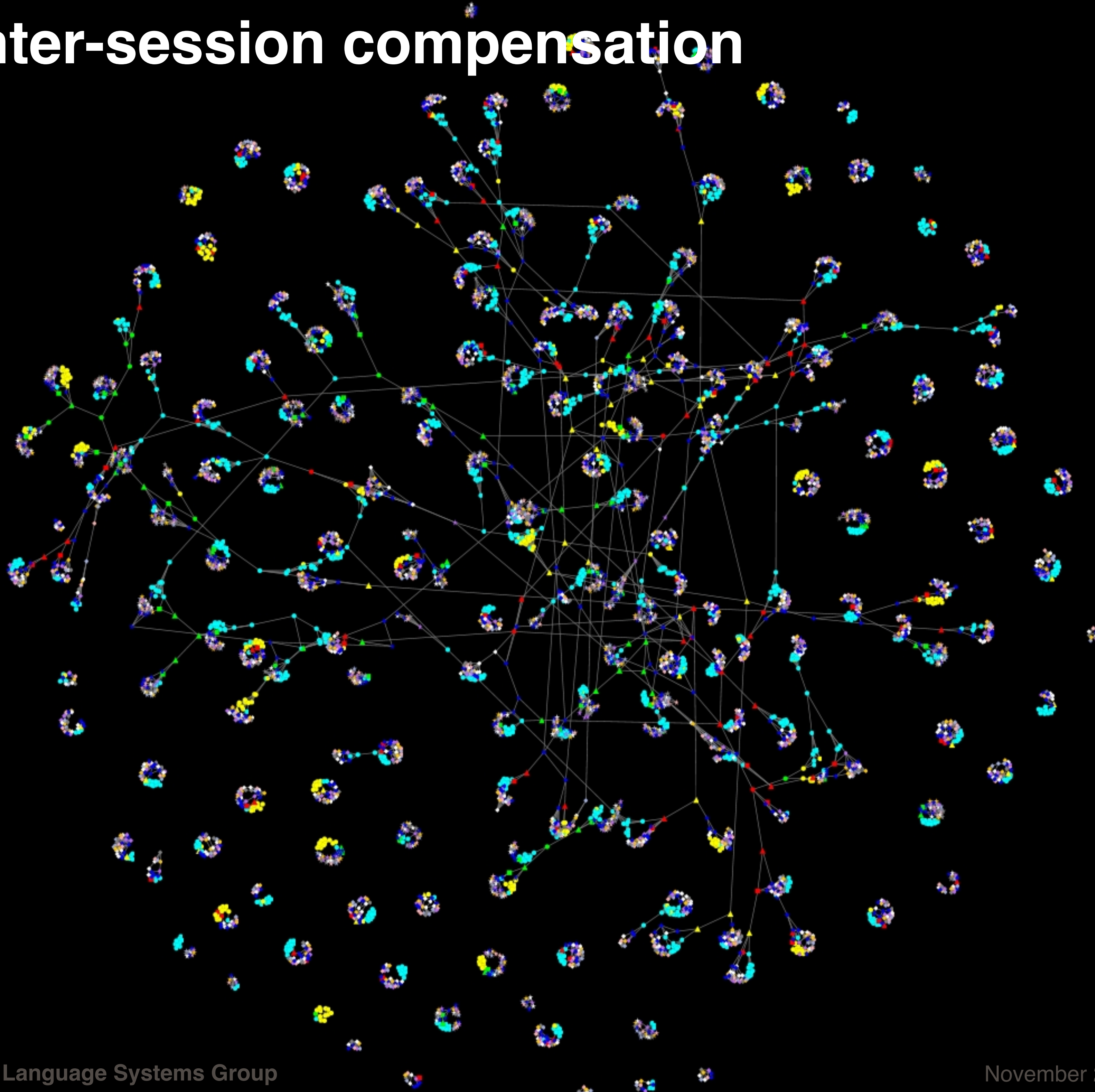
Cell phone
Landline
215573qqn
215573now
Mic_CH08
Mic_CH04
Mic_CH12
Mic_CH13
Mic_CH02
Mic_CH07
Mic_CH05
▲= high VE
■= low VE
●= normal VE
♦=room LDC
* =room HIVE

Without inter-session compensation



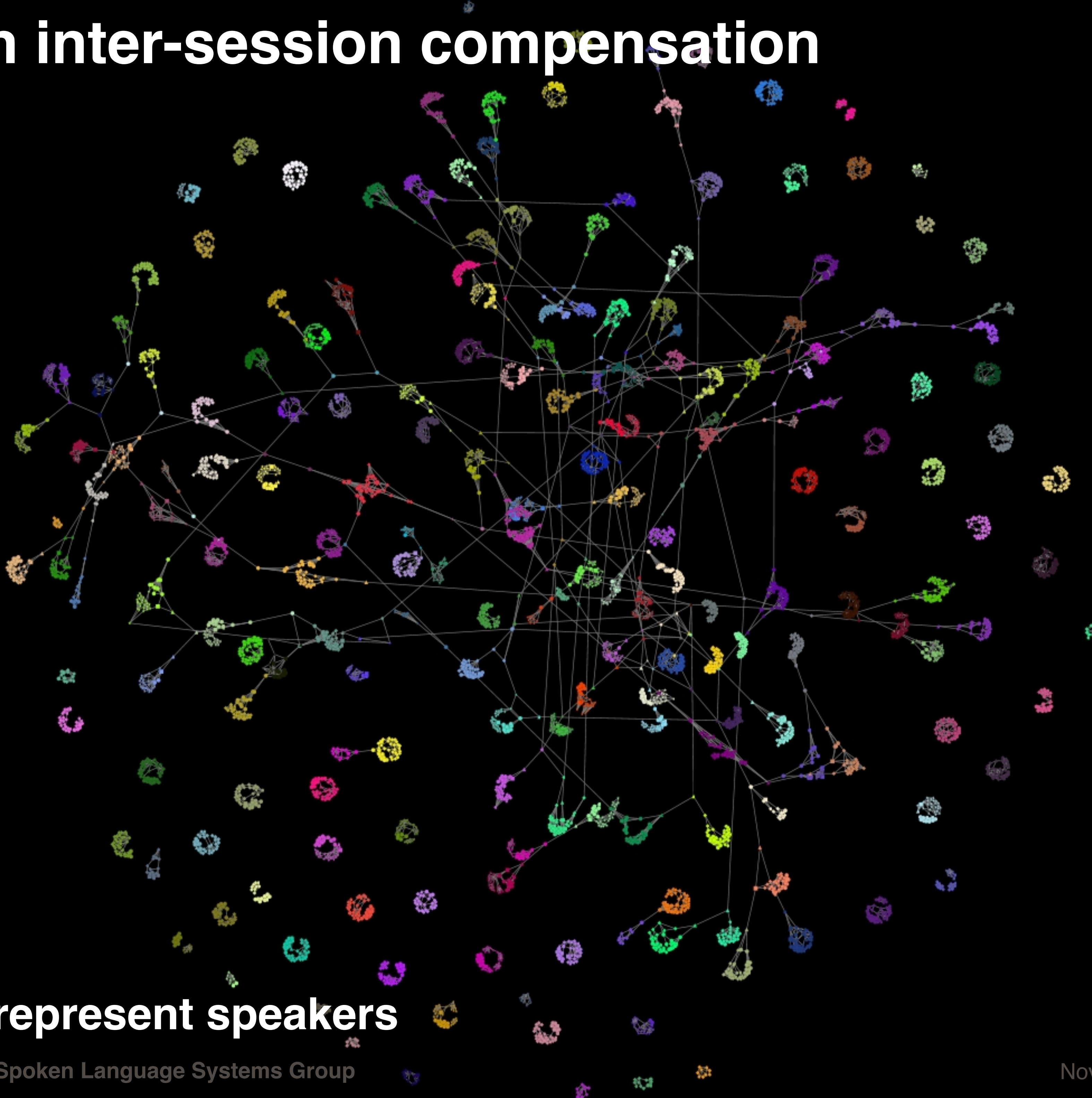
Cell phone
Landline
215573qqn
215573now
Mic_CH08
Mic_CH04
Mic_CH12
Mic_CH13
Mic_CH02
Mic_CH07
Mic_CH05
▲= high VE
■= low VE
●= normal VE
♦=room LDC
* =room HIVE

With inter-session compensation



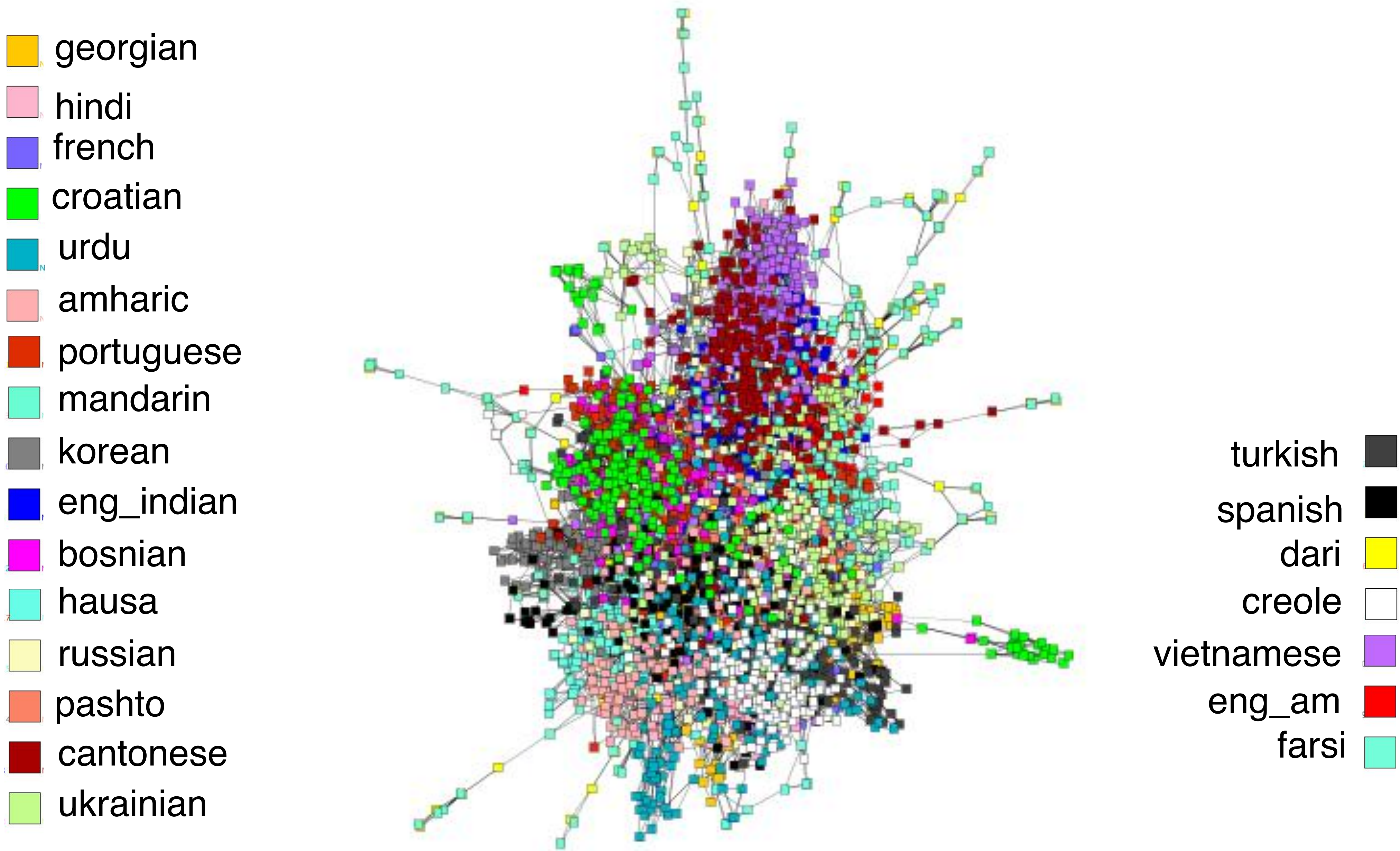
Cell phone
Landline
215573qqn
215573now
Mic_CH08
Mic_CH04
Mic_CH12
Mic_CH13
Mic_CH02
Mic_CH07
Mic_CH05
▲= high VE
■= low VE
●= normal VE
♦=room LDC
* =room HIVE

With inter-session compensation

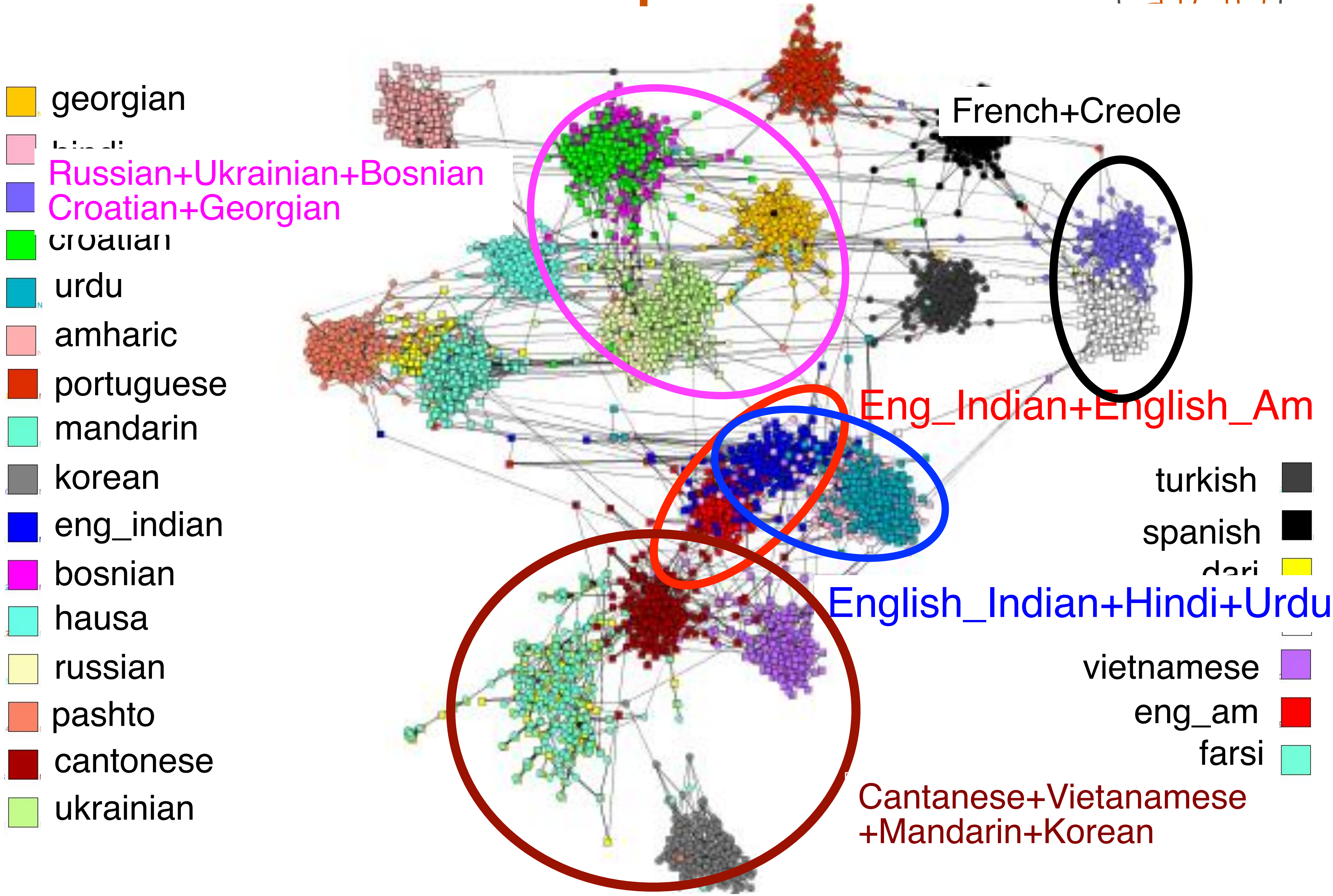


Colors represent speakers

Without inter-session compensation



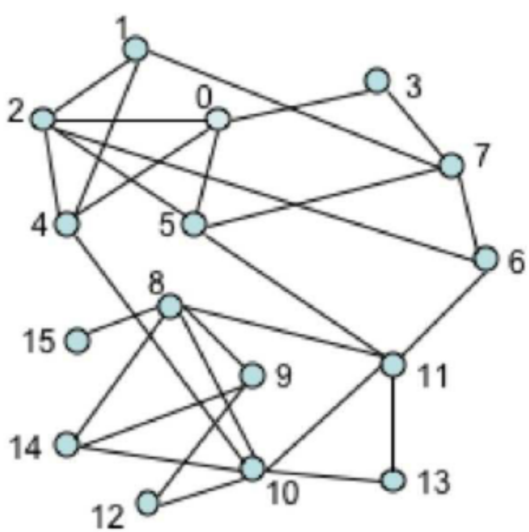
With inter-session compensation



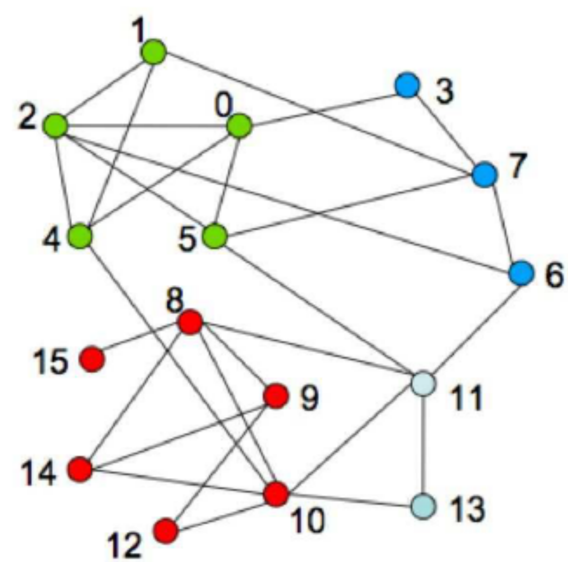
What's next?

- We can build graphs according to certain specifications (i.e., K-NN) and apply the *known* node labels to produce effective and compelling visualizations.

- What can we do with arbitrary graphs with <u>no</u> known labels?



"Big Data"

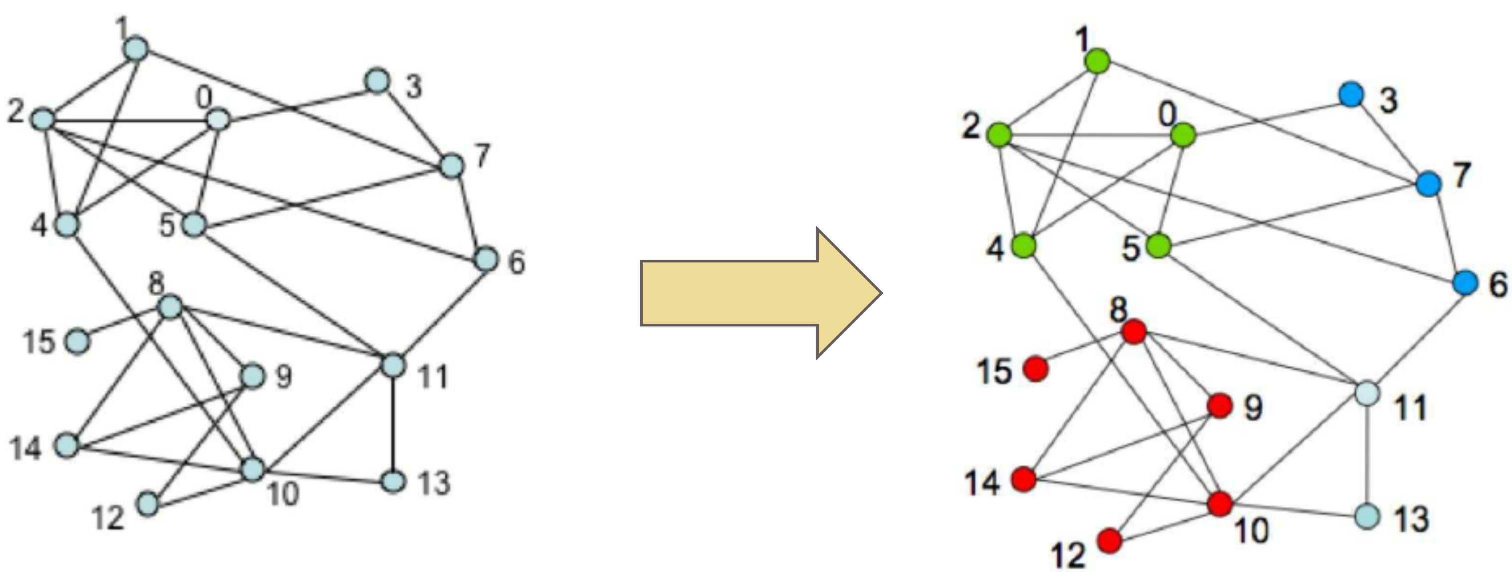


Community Detection

So far,

CSAIL

- Little previous work exists in the speaker recognition field
- Initial and exploratory work presented at ICASSP 2013
- Applied this work to "domain adaptation" over the summer

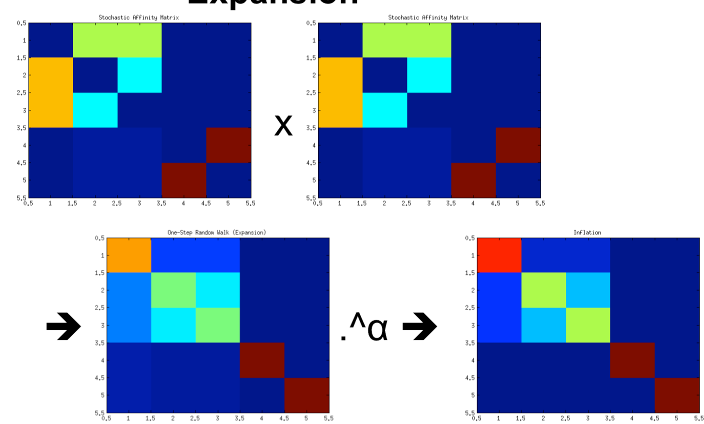


Speaker content graphs **Clustering**

Quick Summary

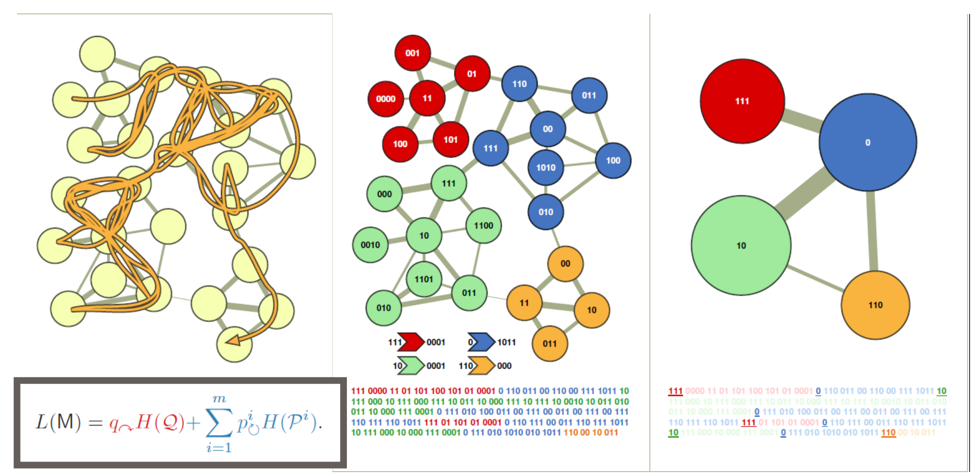
- **Two datasets, ~11,000 utterances each, from NIST SRE's**
- **Different graph constructions**
  - 2-, 5-, 10-, 25-, 50-, 100-NN graphs
    - * **Experimented with "local node-level pruning"**

"Expansion"



"Inflation"

- **Graph clustering algorithms**
  - Agglomerative hierarchical clustering (AHC)
  - Markov Clustering (MCL)
    - * **van Dongen, 2000**
  - Infomap
    - * **Rosvall and Bergstrom, 2008**



Main Takeaways

- **Given an unlabeled speaker content graph, we can do a reasonable job of discovering the underlying speakers.**
- **Agglomerative hierarchical clustering does the best**
  - Need to specifying stopping criterion (i.e., number of speakers)
- **Random-walk methods also do well**
  - Provide reasonable estimates of the number of speakers
  - More dependent on graph-construction parameters

Unsupervised Clustering Approaches for Domain Adaptation in Speaker Recognition Systems

CSAIL

Unsupervised Clustering Approaches for Domain Adaptation in Speaker Recognition Systems

Domain Adaptation & Transfer Learning

- Most current statistical learning techniques assume (incorrectly) that the training and test data come from the same underlying distribution.
- Labeled data may exist in one domain, but we want a model that can also perform well on a related, but not identical, domain.
- Hand-labeling data in a new domain is hard and expensive.
- Can we leverage the original, labeled, “out-of-domain” data when building a model to work on the new, unlabeled, “in-domain data?

Unsupervised Clustering Approaches for Domain Adaptation in Speaker Recognition Systems

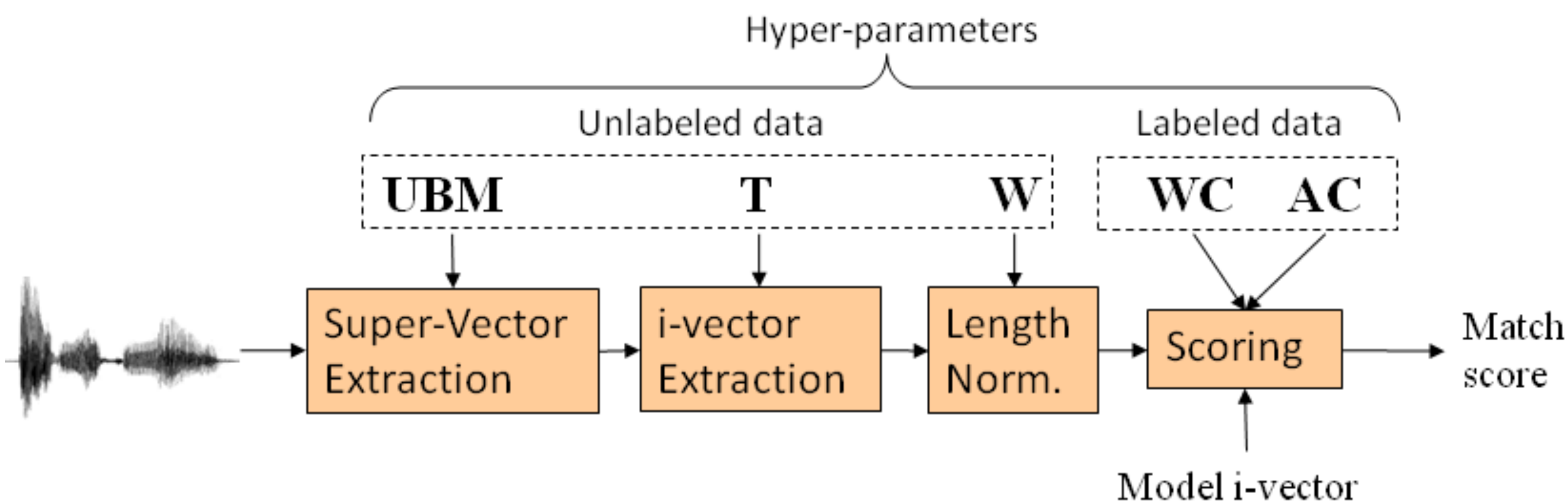
In the context of speaker recognition

- **Current success of i-vector approach has depended upon having access to large amounts of matched and labeled training data**
  - 1000’s of speakers making 10’s of calls
  - Recent SRE’s have become a bit of a data-engineering exercise

- **We can’t realistically expect that most applications will have access to such a large set of labeled data from matched conditions.**

- **How can we design a task to focus research efforts on how to use unlabeled data for adapting system hyper-parameters to a new domain?**

Usage of data (labeled & unlabeled) in an i-vector system



Demonstrating Mismatch

- **Enroll and score**
  - SRE10 telephone speech
    - **Annual/Biannual NIST Speaker Recognition Evaluation (SRE)**

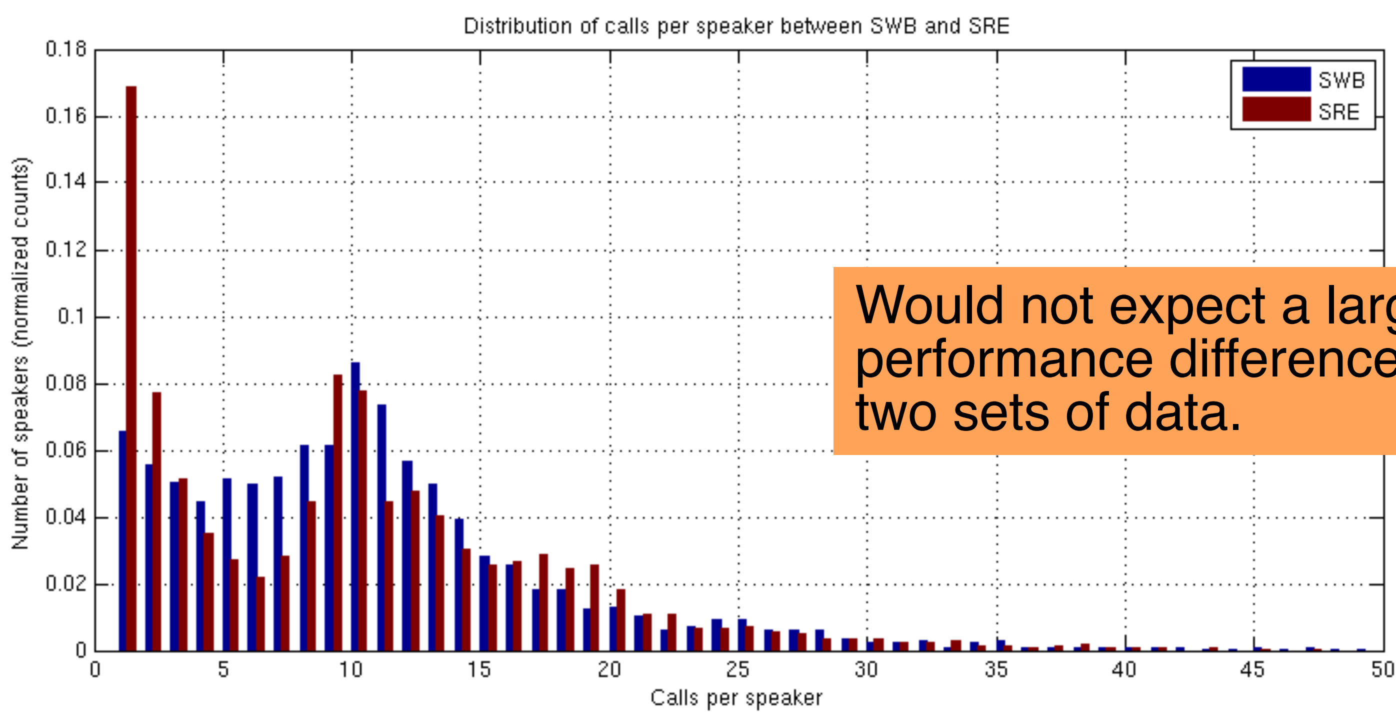
- **Matched, "in-domain" SRE data**
  - All calls from all speakers from SRE 04, 05, 06, and 08 collections

- **Mismatched "out-of-domain" SWB data**
  - All calls from all speakers from Switchboard-I and Switchboard-II

Demonstrating Mismatch

- **Summary statistics for SRE & SWB lists**

Hyper list	# Spkrs	# Males	# Females	# Calls	Avg # calls/spkr	Avg # phone_num/spkr
SWB	3114	1461	1653	33039	10.6	3.8
SRE	3790	1115	2675	36470	9.6	2.8



Would not expect a large performance difference using these two sets of data.

Demonstrating Mismatch

- **Baseline / Benchmark Results (Equal Error Rate – EER)**

UBM & T	Whitening	WC & AC	JHU	MIT
SWB	SWB	SWB	6.92%	7.57%
SWB	SRE	SWB	5.54%	5.52%
SWB	SRE	SRE	2.30%	2.09%
SRE	SRE	SRE	2.43%	2.48%

- **Focus on gap between using SWB/SRE labels for WC & AC**
  - Continue using SWB for UBM&T and SRE for Whitening

Challenge Task Rules

- **Allowed to use SWB data *and* their labels**
- **Allowed to use SRE data but not their labels**
- **Evaluate on SRE10.**

Unsupervised Clustering Approaches

for Domain Adaptation in

Speaker Recognition Systems

Stephen Shum — Spoken Language Systems Group

November 2013

Proposed Framework

- Begin with Σ_{SWB} (WC) and Φ_{SWB} (AC).
- Use PLDA and Σ_{SWB} , Φ_{SWB} to compute pairwise affinity matrix, A, on SRE data.
- Cluster A to obtain hypothesized speaker labels.
- Use labels to obtain Σ_{SRE} and Φ_{SRE}
- Linearly interpolate (via α_{WC} and α_{AC}) between prior (SWB) and new (SRE) covariance matrices to obtain final hyper-parameters:

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

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

- Iterate?

(Unsupervised) Clustering

- **Agglomerative hierarchical clustering (AHC)**
  - Provide the number of clusters at which to stop
- **Graph-based random walk algorithms**
  - Infomap
  - Markov Clustering (MCL)

Initial Results (1000 SRE speakers)

#		# Spkrs K	# Clstrs $\hat{K}$	Clustering Performance Confusion	Purity	Frag.	α^* EER (%) Perfect	Hyp.	Gap	$\alpha = \mathbf{1}$ EER (%) Perfect	Hyp.	Gap
1	AHC	1000	1000*	7.4%	94.9%	1.20	2.37	**2.55**	**7.8%**	2.77	**3.16**	**14%**
2	Infomap	—	918	18.2%	85.9%	1.44	—	2.71	14%	—	3.45	25%
3	MCL	—	997	15.1%	90.3%	1.45	—	2.68	13%	—	3.40	23%

- **α^***
  - Assumes the selection of optimal interpolation parameters (oracle)
- **$\alpha = 1$**
  - Use only the hyper-parameters obtained from hypothesized cluster labels

- **Better clustering → better recognition performance**
  - However, effect is severely attenuated both in recognition results and in the presence of hyper-parameter interpolation!

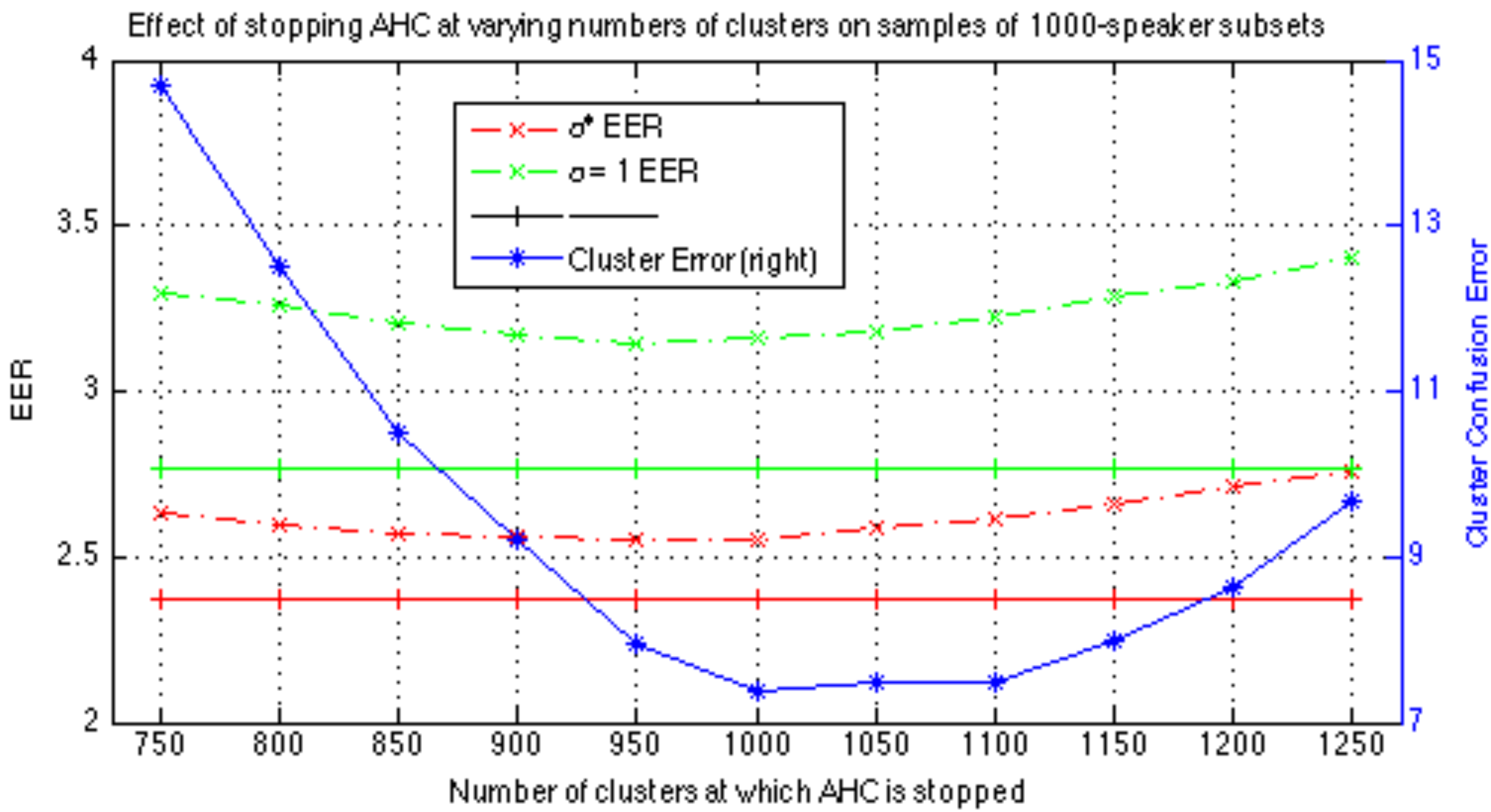
Initial Results (1000 SRE speakers)

#		# Spkrs K	# Clstrs $\hat{K}$	Clustering Performance			α^* EER (%)			$\alpha = \mathbf{1}$ EER (%)		
				Confusion	Purity	Frag.	Perfect	Hyp.	Gap	Perfect	Hyp.	Gap
1	AHC	1000	1000*	7.4%	94.9%	1.20	2.37	**2.55**	**7.8%**	2.77	**3.16**	**14%**
2	Infomap	—	918	18.2%	85.9%	1.44	—	2.71	14%	—	3.45	25%
3	MCL	—	997	15.1%	90.3%	1.45	—	2.68	13%	—	3.40	23%

- **AHC provides best clustering and recognition**
  - Requires number of speakers as stopping criterion
- **Infomap and MCL provide reasonable estimates of #spkrs**
  - Assuming appropriate choice of sparse graph

→ Use Infomap/MCL to estimate #spkrs for AHC

Effect of stopping AHC at different cluster numbers



Initial Results (1000 SRE speakers)

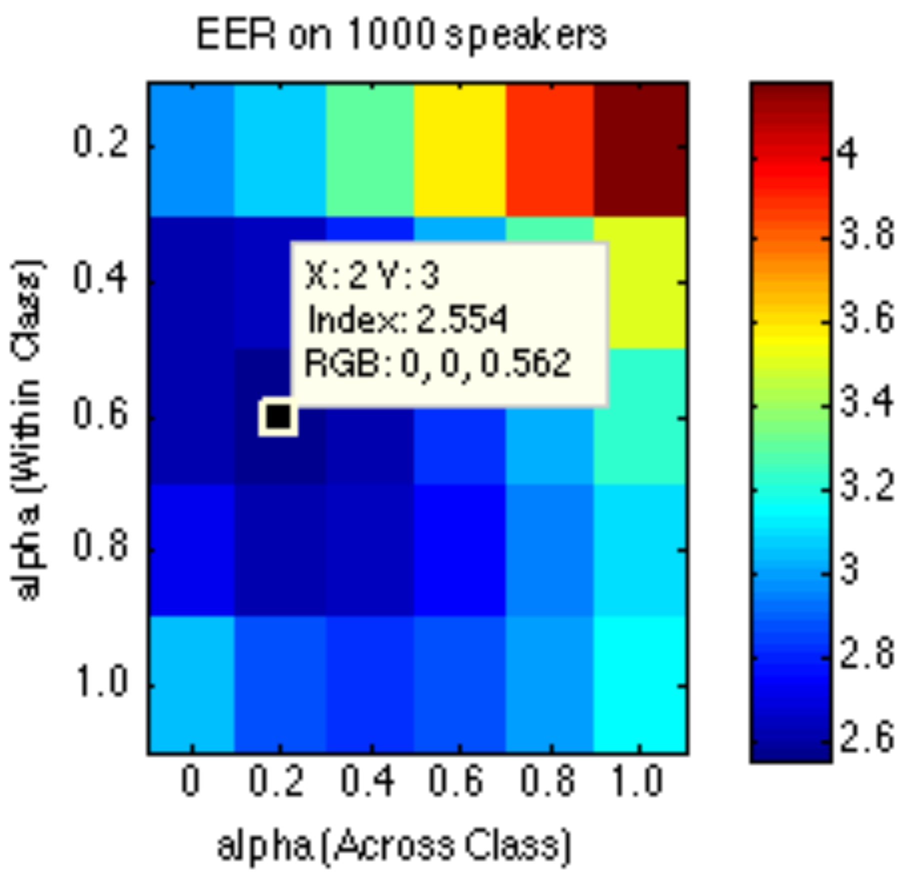
#		# Spkrs K	# Clstrs $\hat{K}$	Clustering Performance Confusion	Purity	Frag.	α^* EER (%) Perfect	Hyp.	Gap	$\alpha = 1$ EER (%) Perfect	Hyp.	Gap
1	AHC	1000	1000*	7.4%	94.9%	1.20	2.37	**2.55**	**7.8%**	2.77	**3.16**	**14%**
2	Infomap	—	918	18.2%	85.9%	1.44	—	2.71	14%	—	3.45	25%
3	MCL	—	997	15.1%	90.3%	1.45	—	2.68	13%	—	3.40	23%
4												
5	Infomap+AHC	1000	918	9.0%	92.6%	1.19	2.37	2.56	8.2%	2.77	3.18	15%
6	MCL+AHC	—	997	7.5%	94.9%	1.20	—	**2.56**	**8.0%**	—	**3.16**	**14%**

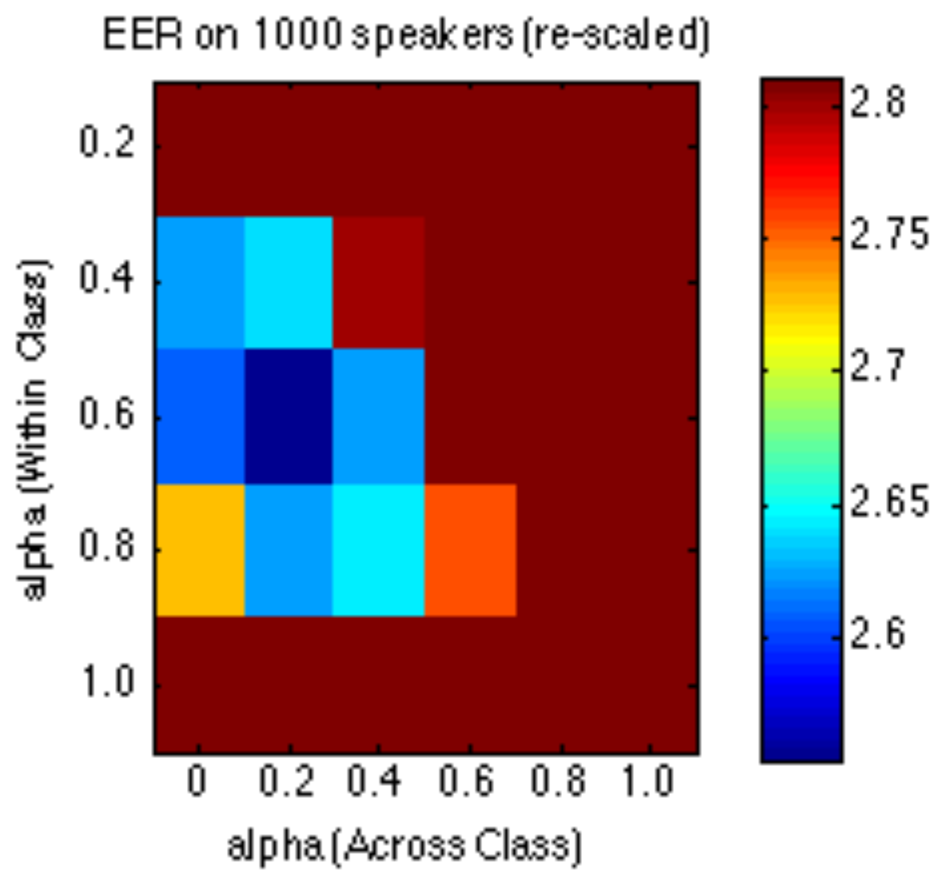
- **AHC provides best clustering and recognition**
  - Requires number of speakers as stopping criterion
- **Infomap and MCL provide reasonable estimates of the number of speakers**
  - Assuming appropriate choice of sparse graph

→ Use Infomap/MCL to estimate #spkrs for AHC

Automatic estimation of α*

- **Still an open and unsolved problem**





Results So Far

- **Via clustering and optimal adaptation**

	$\hat{K}$	Perfect	Hypothesized	Gap (%)
AHC	3790*	2.23	2.58	16%
Infomap+AHC	3196	—	**2.53**	**13%**
MCL+AHC	3971	—	2.61	17%

- **Initial baseline and benchmark**

UBM & T	Whitening	WC & AC	JHU
SWB	SRE	SWB	5.54%
SWB	SRE	SRE	2.30%

Take-home Ideas

- **In the presence of adaptation, α, an imprecise estimate of the number of clusters is forgivable.**
- **A range of adaptation parameters yield decent results.**
  - The selection of optimal values is still an open question.
- **Best automatic system so far obtains SRE10 performance that is within 15% of a system that has access to all speaker labels.**

What's next?

- **Telephone – Telephone domain mismatch**
  - Simple solutions work well already
  - Explicitly identifying the source of the performance degradation
    - **Work currently ongoing**
- **Telephone – Microphone domain mismatch**
  - Expected to be a more difficult problem
    - **Initial experiments pending**
- **Out-of-domain detection**
  - Instance of the canonical outlier/novelty detection problem

Final Words

- **Vector-based representations of speech for speaker and language recognition**
  - UBM-MAP → supervector → i-vector
  - Independent of speech duration
  - Can easily apply known methods for channel/session compensation
- **Graph-based representation of audio databases enables fast and large-scale processing of existing and incoming data**
  - Query-by-example, speaker indexing/clustering, general insights
- **Discussed the application of both ideas in the context of domain adaptation for speaker recognition.**
  - Still a lot to do and learn!